FINDING AND CERTIFYING NUMERICAL ROOTS OF SYSTEMS OF EQUATIONS

A Dissertation
Presented to
The Academic Faculty

By

Kisun Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology

May 2020

FINDING AND CERTIFYING NUMERICAL ROOTS OF SYSTEMS OF EQUATIONS

Approved by:

Dr. Anton Leykin, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Greg Blekherman
School of Mathematics
Georgia Institute of Technology

Dr. Luca Dieci
School of Mathematics
Georgia Institute of Technology

Dr. Josephine Yu
School of Mathematics
Georgia Institute of Technology

Dr. Michael Burr
School of Mathematical Sciences
Clemson University

Date Approved: April 9, 2020

For all the hardships, and for all the will to overcome them.

ACKNOWLEDGEMENTS

First of all, I thank my advisor Anton Leykin for his support, patience and guidance for the last five years; for introducing interesting research topics, for having lots of meetings and discussions, for encouraging me to be an independent mathematician and for sharing his experience always with open mind.

I also would like to thank to my thesis committee, Greg Blekherman, Michael Burr, Luca Dieci and Josephine Yu for assisting me to complete the step toward academic thesis. All of them supported me not only as my thesis committee members, but also teachers and kind colleagues throughout my time in Georgia Tech.

Thanks to nonlinear algebra friends in Georgia Tech, Justin Chen, Timothy Duff, Marc Härkönen, Cvetelina Hill and Bo Lin. Throughout the life in Georgia Tech, they were my supporters, teachers, coworkers, and friends. I was always happy to learn with them, learn from them, work and spend time together both mathematically and in everyday life.

I am also grateful to my friends in Georgia Tech, Jose Acevedo, Skye Binegar, May Cai, Kathy Chen, Christina Giannitsi, Jaewoo Jung, Justin Lanier, Hyunki Min, Jaemin Park, Adrian Perez Bustamante, Thomas Rodewald, Jieun Seong, Kevin Shu and many others who supported my academic path.

Finally, I send gratitude to my family, father, mother and brother for their limitless support, deep love and attention.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis studies numerical analytic aspects of algebraic geometry. Numerical algebraic geometry provides methods to approximate solutions for systems of equations which is computationally expensive for symbolic computations. Since these methods rely on heuristic algorithms, certifying the correctness of their outputs is necessary.

The main focus of the thesis is finding and certifying roots of a given system of equations. Chapter 1 introduces several schemes to approximate roots of a system. Newton's method and homotopy continuation are introduced as core approaches for numerical root approximation. Especially, based on work in [Duf+19], we explain how homotopy continuation and monodromy action are applied to solve parametrized polynomial systems. In Chapters 2 and 3, we study the method of certifying roots obtained by methods in the previous chapter. More precisely, for a given approximation of a solution for a square system of equations, we want to obtain a region which contains the exact solution uniquely. Chapter 2 introduces algorithms to certify regular roots of systems. Methods mainly used are the Krawczyk method and Smale's α-theory, both are based on Newton iteration. For the case of polynomial systems, both methods are implemented in `Macaulay2` in the package `NumericalCertification` [Lee19]. On the other hand, for the case of systems with analytic functions, the Krawczyk method and α-theory can be extended for analytic systems when the oracles from analytic functions are given. These oracles exist for D-finite functions, so that both methods are applied for certifying solutions of systems with D-finite functions. This is based on work in [BLL19]. Chapter 3 tackles the case of multiple roots of systems of equations based on work in [LLZ20]. We introduce the local dual space as a tool to analyze the multiplicity structure of multiple roots. As a method to reduce the singularity of multiple roots, we use the iterative deflation method. Instead of the usual Newton iteration, which requires invertible Jacobian of a system, we find a modified linear operator for Newton iteration which is invertible for a multiple root whose deflation

process terminates after only one iteration. From the modified Newton iteration, the local separation bounds for the multiple root are obtained. Finally, we establish how the local separation bounds can be applied to certify multiple roots of systems of equations.

CHAPTER 1

FINDING NUMERICAL SOLUTIONS OF SYSTEMS OF EQUATIONS

We review methods for finding approximations for solutions of systems of equations. In other words, for a given system of equations $F : \mathbb{C}^n \to \mathbb{C}^m$ with its exact root x^*, i.e. $F(x^*) = 0$, we want to find $x \in \mathbb{C}^n$ such that x is close enough to x^* and the value of $\|F(x)\|$ is small enough.

1.1 Newton's method

Newton's method is one of the most commonly used method for approximating solutions of systems of equations. The idea is to refine an approximation to a solution using a tangent at a given point. For a square system of equations $F : \mathbb{C}^n \to \mathbb{C}^n$, we define *Newton's operator* $N_F(x) := x - F'(x)^{-1}F(x)$. Newton's operator can be applied multiple times, and we define its k-th iteration $N_F^k(x)$. If the sequence $\{N_F^k(x)\}_k$ converges to an exact root x^* of F, then it *converges quadratically* in that we have

$$\lim_{k\to\infty} \frac{\|N_F^{k+1}(x) - x^*\|}{\|N_F^k(x) - x^*\|^2} < M$$

for some $M > 0$. Even though Newton's method provides fast convergence to a root of a system of equations, there are some cases where the method doesn't work. For example, Newton's method cannot be applied for a multiple root as the Jacobian is close to singular around a multiple root. We will discuss the scheme to recover the quadratic convergence of Newton iteration around a multiple root in §3.1.2. Also, Newton's method does not necessarily terminate and it depends on the starting point of the algorithm. Because of these caveats, we sometimes modify Newton's operator or manipulate the given system into an applicable one.

1.2 Gröbner bases

As one of its well-known applications, a Gröbner basis can be used to find solutions for polynomial systems. We first review the definition of a Gröbner basis.

Definition 1. For any given polynomial ring $\mathbb{C}[x_1, \ldots, x_n]$ with its monomial order and a finitely generated ideal $I = \langle f_1, \ldots, f_m \rangle \subset \mathbb{C}[x_1, \ldots, x_n]$, we define a *Gröbner basis* of I as a set $G \subset \mathbb{C}[x_1, \ldots, x_n]$ satisfying that $I = \langle G \rangle$ and $\text{in}(I) = \langle \text{in}(G) \rangle$ where $\text{in}(G) = \{\text{leading term}(g) \mid g \in G\}$.

The following illustrative example from [CLO13] explains how the Gröbner basis technique can be used for finding roots of polynomial systems.

Example 2. [CLO13, §2.8 Example 2] Consider the polynomial system

$$F = \begin{bmatrix} x^2 + y^2 + z^2 - 1 \\ x^2 + z^2 - y \\ x - z \end{bmatrix}$$

and an ideal $I = \langle x^2 + y^2 + z^2 - 1, x^2 + z^2 - y, x - z \rangle$. If we compute a Gröbner basis of I with respect to the lexicographic order, we have the following basis:

$$\begin{aligned} g_1 &= x - z \\ g_2 &= y - 2z^2 \\ g_3 &= z^4 + \tfrac{1}{2}z^2 - \tfrac{1}{4} \end{aligned} \quad .$$

Since g_3 is a univariate polynomial, we solve g_3 using the quadratic formula. Then, we can substitute the obtained values into g_1 and g_2 to eliminate the variable z. It will give roots of the system F. Of course, if finding exact roots for g_3 is not available, then we use Newton's method introduced above to g_3 to find their approximations.

A Gröbner basis is obtained by symbolic algorithms. This means that it often involves

costly computation especially when one deals with a polynomial system with large size. Solving a system of equations using numerical approaches, on the other hand, involves approximates that may produce results faster.

1.3 Homotopy continuation

We review the homotopy continuation algorithm for solving systems of equations. Consider a square polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$ such that the number of solutions for F is finite. The main idea for the homotopy continuation method is to construct a homotopy between the system we would like to solve and the system which is easy to solve. For this task, we create a *start system* G such that solutions of G can be obtained readily and paths connecting solutions of G and F are smooth. Then, we consider a *linear homotopy continuation*

$$H(t) = tG + (1-t)F, \quad t \in [0,1]$$

between the two systems. In an actual implementation, one may use the *predictor-corrector* technique in [SW05] to track the homotopy numerically. When we have an approximation x_0 for a solution $x(t_0)$ of $H(t_0)$ for some $t_0 \in [0,1]$, for $t_1 > t_0$, the predictor step makes a rough approximation x_1 for $x(t_1)$ and then the corrector step refines x_1. For the predictor step, we obtain an approximation x_1 by solving the system of ordinary differential equations

$$\left(\frac{\partial H}{\partial x}x'(t) + \frac{\partial H}{\partial t}\right)_{x=x(t)} = 0$$

which is obtained by differentiating $H(x(t))$ with respect to t. The corrector step is done by Newton's method, introduced in §1.1. Both methods can be performed numerically, and so homotopy continuation gives an approximation for a solution for square polynomial systems.

The homotopy continuation is implemented in several software packages, for example, `NumericalAlgebraicGeometry` [Ley11], `HomotopyContinuation.jl` [BT18],

HOM4PS2 [LLT08], PHCpack [Ver99] and so on. `NumericalAlgebraicGeometry` was mainly used in this thesis.

1.3.1 Cauchy endgame

In this section, we review some additional considerations when we have a singular root on a homotopy path based on arguments from [SW05]. For a randomly chosen $\gamma \in \mathbb{C}$, consider the homotopy

$$H(t) = t\gamma G + (1-t)F, \quad t \in [0,1] \tag{1.1}$$

from 1 to 0. The genericity of γ makes almost sure that the points satisfying the homotopy $H(t)$ are regular for all $t \in (0,1]$, and this is called γ-*trick*. In this section, we review the way to approximate singular solutions at the target point $t = 0$. As we mentioned in §1.1, Newton's method is not applicable at a multiple root. Therefore, we use another method which is called the *Cauchy endgame*.

Consider a Puiseux series $x(t) = \sum_{i=i_0}^{\infty} c_i t^{\frac{i}{m}} \in \mathbb{C}\{\{t\}\}$ with integers i's, a positive integer m and complex numbers c_i. We say that $x(t)$ is *convergent* at the origin if it is absolutely convergent on a punctured disk centered at the origin. The *Newton-Puiseux Theorem* (c.f. [Ked01]) says that a polynomial $f \in \mathbb{C}[x][t]$ has a solution $x(t)$ which is a convergent Puiseux series. Hence, for a homotopy $H(t)$ in (1.1) and its isolated solution $x(0)$ for $H(0)$, we can write a homotopy path $x(t)$ as a Puiseux series

$$x(t) = x(0) + \sum_{i=1}^{\infty} c_i t^{\frac{i}{m}}.$$

The smallest such $m > 0$ is called the *winding number* of $x(t)$ at $t = 0$. Then, the value of $x(0)$ can be computed numerically by integrating the integral

$$\frac{1}{2\pi i} \oint_{\partial D(0,\epsilon)} \frac{\hat{x}(\hat{t})}{\hat{t}} d\hat{t}$$

where $x(t) = \hat{x}(t^{\frac{1}{m}})$, $\hat{t} = t^{\frac{1}{m}}$ and $D(0, \epsilon)$ is a disk with radius ϵ contained in the domain of $\hat{x}(\hat{t})$.

We illustrate the homotopy continuation method for solving polynomial systems with the following example:

Example 3 (Running Example)**.** Let us consider a square polynomial system

$$F(x, y, z) = \begin{bmatrix} x^3 - 3x^2y + 3xy^2 - y^3 - z^2 \\ z^3 - 3z^2x + 3zx^2 - x^3 - y^2 \\ y^3 - 3y^2z + 3yz^2 - z^3 - x^2 \end{bmatrix}$$

which is suggested in [Stu02]. It is known that F has seven roots. Six of them are cyclic shifts of

$$(x, y, z) = (-.14233 \mp .35878i, .14233 \mp .35878i, \pm.15188i)$$

and the last one is at the origin with multiplicity 8.

We solve F using `NumericalAlgebraicGeometry` and it gives total 14 numerical approximations of F. Throughout the thesis, we use this example as a running example and show the correctnesss of these approximations using various methods.

1.4 Solving polynomial systems using monodromy

In this section we introduce the method for solving a generic system in a family of polynomial systems with parametric coefficients using homotopy continuation (§1.3) and monodromy. In other words, we solve a generic system in a family of systems

$$F_p = (f_p^{(1)}, \dots, f_p^{(N)}) = 0, \quad f_p^{(i)} \in \mathbb{C}[p][x],\ i = 1, \dots, N,$$

with finitely many parameters p and n variables x. For the sake of simplicity, we restrict our attention to *linear parametric* families of systems, defined as systems with affine linear

parametric coefficients, such that for a generic p we have a nonempty finite set of solutions x to $F_p(x) = 0$. This implies $N \geq n$. The number of parameters is arbitrary, but we require that for a generic x there exists p with $F_p(x) = 0$.

1.4.1 Basic settings, preliminaries and framework overview

Let $m, n \in \mathbb{N}$. For a point p in a parameter space $\mathbb{C}^m$, we consider a linear space of square polynomial systems F_p of size n where polynomials $f_p^{(1)}, \ldots, f_p^{(n)}$ in F_p share the same monomial support in the fixed variables $x = (x_1, \ldots, x_n)$. We denote a parametrized linear variety of systems by B which is called a *base space*. If we consider an affine linear map $\varphi : p \mapsto F_p$ for $p \in \mathbb{C}^m$, then we have $\varphi(\mathbb{C}^m) = B$.

We consider a set

$$V = \{(F_p, x) \in B \times \mathbb{C}^n \mid F_p(x) = 0\}$$

which is called a *solution variety* and a projection map π from V to B. For a technical assumption, we have the fibre $\pi^{-1}(F_p)$ has only finitely many points for a generic p (this structure is called a *branched covering*). The set of systems in B with non-generic fibre is called the *branch locus* of π.

In this setting, we consider a set $\pi_1(B \setminus D)$ of loops in $B \setminus D$ in the same homotopy equivalence class. A loop in $\pi_1(B \setminus D)$ induces a permutation on the fibre $\pi^{-1}(F_p)$ which is called a *monodromy action*.

Our goal is to find all solutions for one generic system in $\pi(V)$. We first find a pair $(p_0, x_0) \in V$ and use monodromy action to find all other points x satisfying $F_{p_0}(x) = 0$. We further assume that V is irreducible. This assumption implies that the monodromy action is transitive and the converse is also true.

1.4.2 Monodromy

Fix a system $F_p \in B \setminus D$ and consider a loop, i.e. $\tau : [0,1] \to B \setminus D$ and $\tau(0) = \tau(1) = F_p$, avoiding branch points. Suppose that $\pi^{-1}(F_p) = \{x_1, \ldots, x_d\}$, in other words, there are d many points. Then, we can consider a unique lifting $\tilde{\tau}$. The lifting $\tilde{\tau}$ is now a path in V with $\tilde{\tau}(0) = x_i$ and $\tilde{\tau}(1) = x_j$ for some $1 \leq j \leq d$. Note that the reversal of τ and x_j lifts to a reversal of $\tilde{\tau}$. Thus, the loop τ induces a permutation of $\pi^{-1}(F_p)$. In other words, we have a group homomorphism from the fundamental group of $B \setminus D$ based at F_p

$$\varphi : \pi_1(B \setminus D, F_p) \to S_d.$$

The image of φ is called a *monodromy group* associated to $\pi^{-1}(F_p)$. The monodromy group gives an action on the fiber $\pi^{-1}(F_p)$ by permuting its points.

The construction of the monodromy group above holds for an arbitrary covering with finitely many sheets. The monodromy group is a transitive subgroup of S_d whenever the total space is connected. Since we are working over $\mathbb{C}$, this occurs precisely when the solution variety is irreducible.

Remark 4. For a linear family, we can show that there is at most one irreducible component of the solution variety V for which the restriction of the projection $(F_p, x) \mapsto x$ is dominant (that is, its image is dense). We call such component the *dominant component*. Indeed, let U be the locus of points $(F_p, x) \in \pi^{-1}(B \setminus D)$ such that

- the restriction of the x-projection map is locally surjective, and
- the solution to the linear system of equations $F_p(x) = 0$ in p has the generic dimension.

Being locally surjective could be interpreted either in the sense of Zariski topology or as inducing surjection on the tangent spaces. Then either U is empty or $\overline{U}$ is the dominant

component we need, since it is a vector bundle over an irreducible variety, and is hence irreducible.

In the rest of this section, when we say *solution variety*, we mean the *dominant component of the solution variety*. In particular, for sparse systems restricting the attention to the dominant component translates into looking for solutions only in the torus $(\mathbb{C}^*)^n$.

1.4.3 Homotopy continuation

We use homotopy continuation for tracking the path in B. For generic F_{p_1} and F_{p_2} in B, we consider a linear homotopy continuation

$$H(t) = (1-t)F_{p_1} + tF_{p_2}, \quad t \in [0,1]$$

between two systems. If F_{p_1} and F_{p_2} are chosen generically, then we have $H(t)$ outside D. Thus, all systems in $H(t)$ have the same number of finitely many solutions. As this homotopy lies in B, a lifting of $H(t)$ is a path in V and it establishes a one-to-one correspondence between the points in $\pi^{-1}(F_{p_1})$ and $\pi^{-1}(F_{p_2})$. This path in V is called a *homotopy path*.

Remark 5. We use γ-trick introduced in §1.3.1 in order to avoid singular homotopy paths. Note that γF_p for $\gamma \in \mathbb{C} \setminus \{0\}$ has the same solutions as F_p. Let us scale both ends of the homotopy by taking a homotopy between $\gamma_1 F_{p_1}$ and $\gamma_2 F_{p_2}$ for generic γ_1 and γ_2. If the coefficients of F_p are homogeneous in p then

$$H'(t) = (1-t)\gamma_1 F_{p_1} + t\gamma_2 F_{p_2} = F_{(1-t)\gamma_1 p_1 + t\gamma_2 p_2}, \quad t \in [0,1],$$

is a homotopy matching solutions $\pi^{-1}(F_{p_1})$ and $\pi^{-1}(F_{p_2})$ where the matching is potentially different from that given by $H(t)$. Similarly, for an affine linear family, $F_p = F'_p + C$ where F'_p is homogeneous in p and C is a constant system, we have

$$H'(t) = (1-t)\gamma_1 F_{p_1} + t\gamma_2 F_{p_2} = F'_{(1-t)\gamma_1 p_1 + t\gamma_2 p_2} + ((1-t)\gamma_1 + t\gamma_2)C.$$

We ignore the fact that $H'(t)$ may go outside B for $t \in (0,1)$, since its rescaling,

$$\begin{aligned} H''(t) &= \frac{1}{(1-t)\gamma_1 + t\gamma_2} H'(t) \\ &= F'_{\frac{(1-t)\gamma_1 p_1 + t\gamma_2 p_2}{(1-t)\gamma_1 + t\gamma_2}} + C = F_{\frac{(1-t)\gamma_1 p_1 + t\gamma_2 p_2}{(1-t)\gamma_1 + t\gamma_2}}, \quad t \in [0,1], \end{aligned}$$

does not leave B and clearly has the same homotopy paths. Note that $H''(t)$ is well defined as $(1-t)\gamma_1 + t\gamma_2 \neq 0$ for all $t \in [0,1]$ for generic γ_1 and γ_2.

1.4.4 Graph of homotopies: main ideas

To organize the discovery of new solutions we represent the set of homotopies by a finite undirected graph G. Let $E(G)$ and $V(G)$ denote the edge and vertex set of G, respectively. Any vertex v in $V(G)$ is associated to a point F_p in the base space. An edge e in $E(G)$ connecting v_1 and v_2 in $V(G)$ is decorated with two complex numbers, γ_1 and γ_2, and represents the linear homotopy connecting $\gamma_1 F_{p_1}$ and $\gamma_2 F_{p_2}$ along a line segment (Remark 5). We assume that both p_i and γ_i are chosen so that the segments do not intersect the branch locus. Choosing these at random satisfies the assumption almost surely, since the exceptional set of choices where such intersections happen is contained in a real Zariski closed set, see [SW05, Lemma 7.1.3].

We allow multiple edges between two distinct vertices but no loops, since the latter induce trivial homotopies. For a graph G to be potentially useful in a monodromy computation, it must contain a cycle. Some of the general ideas behind the structure of a graph G are listed below.

- For each vertex v_i, we maintain a subset of *known* points $Q_i \subset \pi^{-1}(F_{p_i})$.
- For each edge e between v_i and v_j, we record the two complex numbers γ_1 and γ_2 and we store the known partial correspondences $C_e \subset \pi^{-1}(F_{p_i}) \times \pi^{-1}(F_{p_j})$ between known points Q_i and Q_j.

- At each iteration, we pick an edge and direction, track the corresponding homotopy starting with yet unmatched points, and update known points and correspondences between them.

- We may obtain the initial "knowledge" as a *seed pair* (p_0, x_0) by picking $x_0 \in \mathbb{C}^n$ at random and choosing p_0 to be a generic solution of the linear system $F_p(x_0) = 0$.

We list basic operations that result in transition between one state of our algorithm captured by G, Q_i for $v_i \in V(G)$, and C_e for $e \in E(G)$ to another.

1. For an edge $e = v_i \xleftrightarrow{(\gamma_1,\gamma_2)} v_j$, consider the homotopy

$$H^{(e)} = (1-t)\gamma_1 F_{p_i} + t\gamma_2 F_{p_j}$$

   where $(\gamma_1, \gamma_2) \in \mathbb{C}^2$ is the label of e.

   - Take start points S_i to be a subset of the set of known points Q_i that do not have an established correspondence with points in Q_j.
   - Track S_i along $H^{(e)}$ for $t \in [0, 1]$ to get $S_j \subset \pi^{-1}(F_{p_j})$.
   - Extend the known points for v_j, that is, $Q_j := Q_j \cup S_j$ and record the newly established correspondences.

2. Add a new vertex corresponding to F_p for a generic $p \in B \setminus D$.

3. Add a new edge $e = v_i \xleftrightarrow{(\gamma_1,\gamma_2)} v_j$ between two existing vertices decorated with generic $\gamma_1, \gamma_2 \in \mathbb{C}$.

Example 6. Figure 1.1 shows a graph G with 2 vertices and 3 edges embedded in the base space B with paths partially lifted to the solution variety, which is a covering space with 3 sheets. The two fibers $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$ are connected by 3 partial correspondences induced by the liftings of three edge-paths.

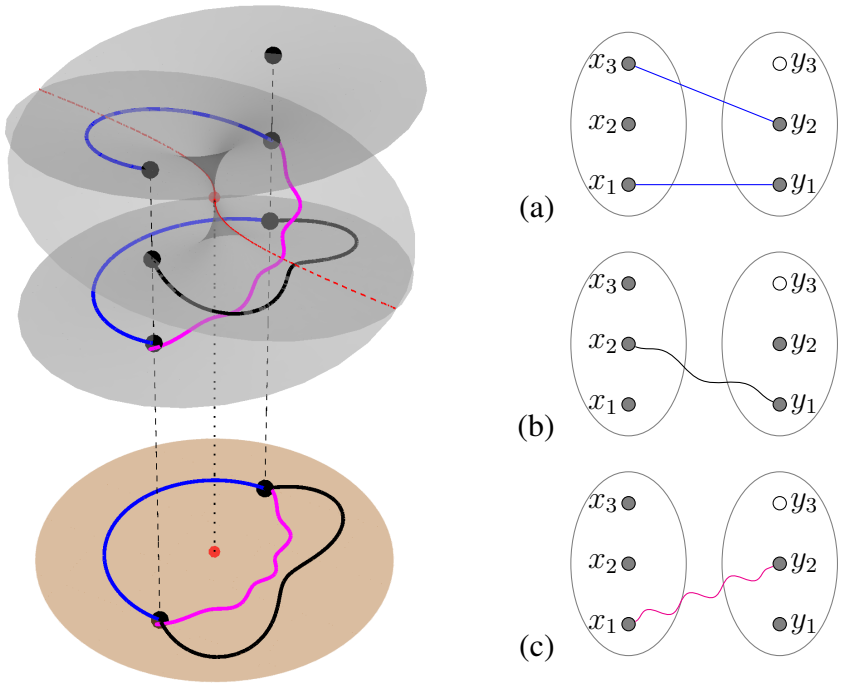

Figure 1.1: Selected liftings of 3 edges connecting the fibers of 2 vertices and induced correspondences.

Note that several aspects in this illustration are fictional. There is only one branch point in the actual complex base space B that we would like the reader to imagine. The visible self-intersections of the solution variety V are an artifact of drawing the picture in the real space. Also, in practice we use homotopy paths as simple as possible, however, here the paths are more involved for the purpose of distinguishing them in print.

An algorithm that we envision may hypothetically take the following steps:

1. *seed* the first fiber with x_1;
2. use a lifting of edge e_a to get y_1 from x_1;
3. use a lifting of edge e_b to get x_2 from y_1;
4. use a lifting of edge e_c to get y_2 from x_1;

5. use a lifting of edge e_a to get x_3 from y_2.

Note that it is *not* necessary to complete the correspondences (a), (b), and (c). Doing so would require tracking 9 continuation paths, while the hypothetical run above uses only 4 paths to find a fiber. The additional considerations for probabilistic analysis and parallelization were done in [Bli+18].

Example 7. Figure 1.2 illustrates two partial correspondences associated to two edges e_a and e_b, both connecting two vertices v_1 and v_2 in $V(G)$. Each vertex v_i stores the array of known points Q_i, which are depicted in solid. Both correspondences in the picture are subsets of a perfect matching, a one-to-one correspondence established by a homotopy associated to the edge.

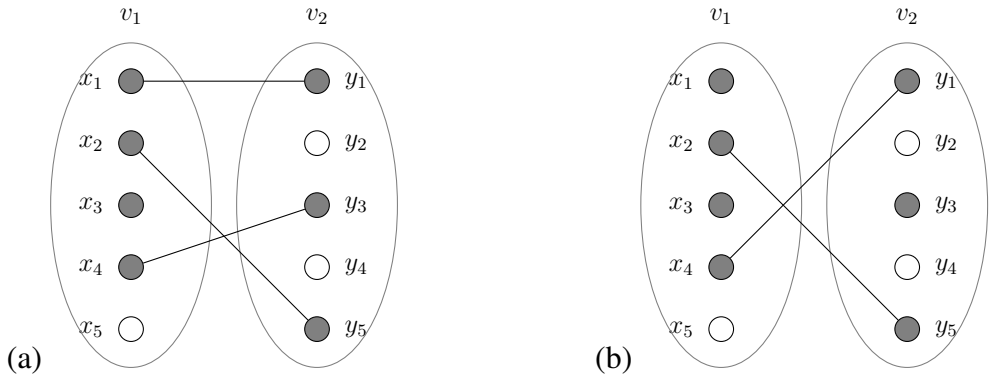

Figure 1.2: Two partial correspondences induced by edges e_a and e_b for the fibers of the covering map of degree $d = 5$ in Example 7.

Note that taking the set of start points $S_1 = \{x_3\}$ and following the homotopy $H^{(e_a)}$ from left to right is *guaranteed* to discover a new point in the second fiber. On the other hand, it is impossible to obtain new knowledge by tracking $H^{(e_a)}$ from right to left. Homotopy $H^{(e_b)}$ has the *potential* to discover new points if tracked in either direction. We can choose $S_1 = \{x_1, x_3\}$ as the starting points for one direction and $S_2 = \{y_3\}$ for the other. In this scenario, following the homotopy from left to right is guaranteed to produce at least one new point, while going the other way may either deliver a new point or augment the

correspondences between the already known points. If the correspondences in (a) and (b) are completed to one-to-one correspondences of the fibers, taking the homotopy induced by the edge e_a from left to right followed by the homotopy induced by edge e_b from right to left would produce a permutation. However, the group generated by this permutation has to stabilize $\{x_2\}$, therefore, it would not act transitively on the fiber of v_1. One could also imagine a completion such that the given edges would not be sufficient to discover x_5 and y_4.

1.5 Subdivision methods

In this section, we briefly point out *subdivision methods* which are also well-known numerical methods for approximating solutions for systems of equations. For a given region for a solution of a system of equations, we subdivide the domain until we know that either there is no solution in all subdivided regions or there is a solution in a piece of subdivided regions and the piece is small enough that all point in this piece are "nice" approximations for the solution.

The *bisection method* is one of commonly used subdivision methods. Combining this with *interval arithmetic* and Newton's method, we have the *Krawczyk method* which will be introduced in §2.2.

CHAPTER 2

CERTIFYING REGULAR SOLUTIONS OF SYSTEMS OF EQUATIONS

In this chapter, we focus on certifying numerical approximations of regular solutions of square systems of equations, i.e. given (a finite description of) a compact region $I \subset \mathbb{C}^n$ (or $\mathbb{R}^n$) that is conjectured to contain a unique root, execute an algorithm which produces a certificate for the existence and uniqueness of a root in I. The algorithms considered in this chapter are based on α-*theory* and the *Krawczyk test*, which, in turn, are based on Newton iteration. We study how these methods work on polynomial systems by their implementations in `Macaulay2` and an example from [Duf+19]. We extend our interest into certifying roots of systems of analytic equations as the theory originally derived for α-theory [Sma86] and the Krawczyk operator [Kra69] applies to arbitrary square systems of analytic functions.

2.1 Setting

We describe the classes of functions by seeding a class with a set of *basic functions* (we informally call them *ingredients*) and then extending these functions by recursively applying the basic arithmetic operations (addition and multiplication) to the basic functions and constants finitely many times. By adding more variables and equations, other operations, such as division and composition, are possible in the construction.

More explicitly, suppose that the basic functions include both the coordinate functions and additional basic functions $\{g_1, \ldots, g_m\}$. Then, the systems of equations that we con-

struct can be written in the following form (after an appropriate change of variables):

$$F(x) := \begin{bmatrix} p_1(x_1, \ldots, x_{n+m}) \\ \vdots \\ p_n(x_1, \ldots, x_{n+m}) \\ x_{n+1} - g_1(x_1) \\ \vdots \\ x_{n+m} - g_m(x_m) \end{bmatrix} \tag{2.1}$$

where $p_i \in \mathbb{C}[x_1, \ldots, x_{n+m}]$ for $i = 1, \ldots, n$.

Suppose that the basic functions are the coordinate functions $\{x_1, \ldots, x_n\}$, then the class of functions is $\mathbb{C}[x] = \mathbb{C}[x_1, \ldots, x_n]$, i.e. the class of polynomial systems of equations. This class appears frequently in geometric problems (e.g. [BLR15]) and can be effectively studied via α-theory and the Krawczyk operator since all but finitely many derivatives vanish. For a practical implementation of the α-theory approach in this setting, see [HS12].

When the basic functions are the coordinate functions along with univariate analytic functions which satisfy linear differential equations with constant coefficients, the resulting class of functions are the polynomial-exponential functions. In [HL17], Hauenstein and Levandovskyy extend α-theory-based certification to this case.

2.2 The Krawczyk method

In this section, we develop the theory of interval arithmetic and the Krawczyk operator, an interval-based generalization of the Newton operator. We explicitly describe the oracles which are necessary so that the theory described in this section can be developed into an algorithm.

2.2.1 Interval arithmetic

Interval arithmetic performs conservative computations with intervals in order to produce certified computations. For example, suppose that $[a, b]$ and $[c, d]$ are isolating intervals for $x, y \in \mathbb{R}$, i.e. $x \in [a, b]$ and $y \in [c, d]$. Then, interval arithmetic formalizes the conclusion that $x + y \in [a + c, b + d]$. More precisely, for any arithmetic operation $\odot$,

$$[a, b] \odot [c, d] = \{x \odot y : x \in [a, b], y \in [c, d]\}.$$

For the standard arithmetic operations, there are formulas for the interval versions of these operators, see, e.g. [MKC09] for more details.

These methods can extend to complex numbers by writing intervals in $\mathbb{C}$ as $[a_1, a_2] + [b_1, b_2]i$. In this case, multiplication of complex interval numbers is computed as

$$\begin{aligned}([a_1, a_2] + [b_1, b_2]i)([c_1, c_2] + [d_1, d_2]i) \\ = ([a_1, a_2][c_1, c_2] - [b_1, b_2][d_1, d_2]) + ([a_1, a_2][d_1, d_2] + [b_1, b_2][c_1, c_2])i. \quad (2.2)\end{aligned}$$

We observe that the image of this product may be strictly larger than the set of possible products of elements from the pair of complex intervals. This formulation, however, is critically important in our development of the Krawczyk method in §2.2.2.

We write $\mathbb{IC}$ for the set of intervals in $\mathbb{C}$, and we write $\mathbb{IC}^n$ for the set of n-dimensional boxes in $\mathbb{C}^n$, i.e. n-fold products of intervals in $\mathbb{C}$. For an open set $U \subseteq \mathbb{C}^n$, we write $\mathbb{I}U$ for intervals in $\mathbb{IC}^n$ which are contained in U. For a function $F : U \to \mathbb{C}$, an oracle interval extension of F is an oracle $\Box F : \mathbb{I}U \to \mathbb{IC}$ such that for any $I \in \mathbb{I}U$,

$$\Box F(I) \supseteq F(I) := \{F(x) : x \in I\}.$$

In other words, $\Box F(I)$ is an interval containing the image of F on I. For polynomial sys-

tems, such oracles can be constructed using interval arithmetic, see, e.g. [MKC09; RR84] for details.

2.2.2 The Krawczyk method

The Krawczyk operator combines both interval arithmetic and a generalization of the Newton operator in order to develop a certified test for an isolated root of a square system of equations in a region. The Krawczyk operator is one member of the family of interval-based Newton-type methods, see, e.g. [MKC09, §8] and the references therein for more details. In most presentations of the Krawczyk operator, see, e.g. [Kra69; MKC09], the operator is only described for real variables. There are some subtle differences that arise in the complex setting; therefore, in this section, we provide the theory for the Krawczyk operator for complex variables.

Suppose that, for an open set $U \subset \mathbb{C}^n$, $F : U \to \mathbb{C}^n$ is a square differentiable system of functions and let $Y \in GL_n$ the set of $n \times n$ invertible matrices. We observe that $F(x^*) = 0$ if and only if x^* is a fixed point of $G(x) := x - YF(x)$. We note that if Y were replaced by $F'(x^*)^{-1}$, then this function would be the Newton operator. The correspondence between the fixed points of G and the roots of F is the motivation for the Krawczyk operator:

Definition 8. Let $U \subset \mathbb{C}^n$ be an open set and $F : U \to \mathbb{C}^n$ be a square differentiable system of functions such that F' has an interval extension $\square F'$. Let $y \in I \in \mathbb{I}U$ and $Y \in GL_n$. The *Krawczyk operator* centered at y is defined to be

$$K_y(I) := y - YF(y) + (I_n - Y\square F'(I))(I - y),$$

where I_n is the n-dimensional identity matrix.

When the domain and codomain are real, the Krawczyk operator is an interval extension of the function G using the *mean value form*, see, e.g. [MKC09, §6]. In the complex case, however, there is no mean value theorem, but with the definition of complex multiplication

for intervals from Equation (2.2), the Krawczyk operator remains an interval extension of the function G.

Lemma 9. Let $U \subset \mathbb{C}^n$ be an open set and $F : U \to \mathbb{C}^n$ be a square differentiable system of functions such that F' has an interval extension $\square F'$. Let $y \in I \in \mathbb{I}U$ and $Y \in GL_n$. Then,

$$G(I) \subseteq K_y(I).$$

Proof. We observe that $K_y(I) = G(y) + (I_n - Y\square F'(I))(I - y)$, so it is enough to show that for any $z \in I$, $G(z) - G(y) \in (I_n - Y\square F'(I))(I - y)$. Let $w = \Re(z) + i\Im(y)$; we note that $w \in I$ since I is a rectangle. Then, we consider the real path form y to w and the purely imaginary path from w to z. Considering these two paths as functions of a real variable, we use the mean value theorem on each path and on the real and imaginary parts of G separately. Fix $1 \leq j \leq n$. After applying the Cauchy-Riemann equations, each of $G_j(w) - G_j(y)$ and $G_j(z) - G_j(w)$ can be written in terms of the real and imaginary parts of G_j' at appropriate points times $(w - y)$ or $(z - w)$. Then, the sum of these two formulae correspond to elements of the four products appearing in Equation (2.2). By repeating this computation for each j, we conclude that $G(z) \in K_y(I)$. We begin by observing that $I_n - Y\square F'(I)$ is an interval matrix containing $G'(I)$. Our plan, for a fixed $z \in I$, is to write $G(z) - G(y)$ in terms of elements of $G'(I)$, $\Re(z - y)$, and $\Im(z - y)$ in order to conclude the desired containment.

Let $w = \Re(z) + i\Im(y)$, and consider the path from y to w, which is a real path, followed by the path from w to z, which is purely imaginary path. Fix $1 \leq j \leq n$. By the real mean value theorem, there are some c_1 and c_2 along the line between y and w so that $\nabla_{\Re}(\Re G_j(c_1)) \cdot (w - y) = \Re G_j(w) - \Re G_j(y)$ and $\nabla_{\Re}(\Im G_j(c_2)) \cdot (w - y) = \Im G_j(w) - \Im G_j(y)$. Here, the subscript indicates that the derivative is only being taken with respect to the real variable. Similarly, along the line between w and z, there are some c_3 and c_4 so that $\nabla_{\Im}(\Re G_j(c_3)) \cdot \Im(z - w) = \Re G_j(z) - \Re G_j(w)$ and $\nabla_{\Im}(\Im G_j(c_4)) \cdot \Im(z - w) =$

$\Im G_j(z) - \Im G_j(w)$, where the derivative is being taken with respect to the complex variable. Putting these together (and multiplying by i as appropriate), we get

$$G_j(w) = G_j(y) + \nabla_{\Re}(\Re G_j(c_1)) \cdot (w-y) + i\nabla_{\Re}(\Im G_j(c_2)) \cdot (w-y)$$
$$G_j(z) = G_j(w) - i\nabla_{\Im}(\Re G_j(c_3)) \cdot (z-w) + \nabla_{\Im}(\Im G_j(c_4)) \cdot (z-w).$$

Using the Cauchy-Riemann equations, we find that

$$G_j(w) = G_j(y) + \Re G_j'(c_1) \cdot (w-y) + i\Im G_j'(c_2) \cdot (w-y)$$
$$G_j(z) = G_j(w) + i\Im G_j'(c_3) \cdot (z-w) + \Re G_j'(c_4) \cdot (z-w).$$

Therefore,

$$G_j(z) = G_j(y) + \Re G_j'(c_1) \cdot (w-y) + i\Im G_j'(c_3) \cdot (z-w) + \Re G_j'(c_4) \cdot (z-w) + i\Im G_j'(c_2) \cdot (w-y). \tag{2.3}$$

Finally, we observe that since each c_i is in I, the real and imaginary parts of $G_j'(c_i)$ are in the j^{th} row of $I_n - Y\Box F'(I)$. In addition, since $w-y = \Re(z-y)$ and $z-w = i\Im(z-y)$, it follows that the differences $w-y$ and $z-w$ are also in the corresponding real and imaginary parts of $I-y$. Finally, the four products appearing in Equation (2.3) correspond to elements of the four products appearing in Equation (2.2). By repeating this for each $1 \leq j \leq n$, independently, we conclude that $G(z) \in K_y(I)$ and the desired inclusion holds. $\square$

In the following theorem, we collect a few facts about detecting the existence and uniqueness of roots using the Krawczyk operator. We include the proof for completeness.

Theorem 10 (c.f. [Kra69])**.** Let $U \subset \mathbb{C}^n$ be an open set and $F : U \to \mathbb{C}^n$ be a square differentiable system of functions such that F' has an oracle interval extension $\Box F'$. Let $y \in I \in \mathbb{I}U$ and $Y \in GL_n$. The following hold:

1. If $x \in I$ is a root of F, then $x \in K_y(I)$,

2. If $K_y(I) \subset I$, then there is a root of F in I, and

3. If I contains a root of F and $\sqrt{2}\|I_n - Y\Box F'(I)\| < 1$, then the root in I is unique. Here, $\|I_n - Y\Box F'(I)\|$ denotes the maximum operator norm of a matrix in $I_n - Y\Box F'(I)$ under the max-norm.

Proof. (1) Since x is a fixed point of the function G if and only if x is a root of F, by the properties of interval extensions, if $x \in I$ is a root of F, then $G(x) = x$ is in $K_y(I)$. (2) If $K_y(I) \subset I$, then the image of the function G on I is a subset of I, so, by Brouwer's fixed point theorem, G has a fixed point, i.e. a root of F. (3) We observe that by expanding the proof of Lemma 9, we find that for all $z_1, z_2 \in I$,

$$G(z_1) - G(z_2) \in \Box G'(I) \cdot \Re(z_1 - z_2) + \Box G'(I) \cdot \Im(z_1 - z_2).$$

Thus,

$$\|G_1(z_1) - G(z_2)\|_\infty \leq \|I_n - Y\Box F'(I)\|\|\Re(z_1 - z_2)\|_\infty + \|I_n - Y\Box F'(I)\|\|\Im(z_1 - z_2)\|_\infty.$$

The Cauchy-Schwartz inequality and the assumption imply that

$$\|G_1(z_1) - G(z_2)\|_\infty \leq \sqrt{2}\|I_n - Y\Box F'(I)\|\|z_1 - z_2\|_\infty < \|z_1 - z_2\|_\infty,$$

and we conclude that the G function is contractive within I. □

Remark 11. The results of Theorem 10 apply when $\mathbb{C}$ is replaced by $\mathbb{R}$. In fact, in the case of $\mathbb{R}$, the uniqueness test simplifies to $\|I_n - Y\Box F'(I)\| < 1$, i.e. without the $\sqrt{2}$ factor.

Theorem 10 serves as a proof of correctness of the following algorithm.

KrawczykTest$(F, I, Y, y, \square F')$**:**

Input: A square differentiable system of functions $F : U \to \mathbb{C}^n$ for an open set $U \subset \mathbb{C}^n$, an interval $I \in \mathbb{I}U$, an invertible matrix $Y \in GL_n$, a point $y \in I$ and an interval extension $\square F'$.

Output: The boolean value of a condition that implies that "the interval I contains a unique nonsingular root x of F".

return $K_y(I) \subset I$ **and** $\sqrt{2}\|I_n - Y\square F'(I)\| < 1$

In practice, the preconditioning matrix Y is chosen to make $\|I_n - Y\square F'(I)\|$ as small as possible. Without additional information, a good choice is often an approximation to $F'(m(I))^{-1}$, provided it exists, along with $y = m(I)$, i.e. the midpoint of I.

We also observe that it might not be possible to evaluate $F(y)$ exactly. Therefore, we consider a generalization of the Krawczyk operator. Suppose that there is an oracle $\square F$ which, on input $y \in \mathbb{C}^n$, returns an interval $\square F(y)$ containing $F(y)$. Then, we may replace $F(y)$ by $\square F(y)$ in the definition of the Krawczyk operator as follows:

$$\square K_y(I) = y - Y\square F(y) + (I_n - Y\square F'(I))(I - y). \tag{2.4}$$

We observe that $K_y(I) \subset \square K_y(I)$. Therefore, when the corresponding existence and uniqueness results hold for $\square K_y(I)$, they also hold for $K_y(I)$. By combining this operator with Theorem 10, we arrive at a certified test for the Krawczyk operator. In particular, checking that both $\square K_y(I) \subset I$ and $\sqrt{2}\|I_n - Y\square F'(I)\| < 1$ hold, we certify that I contains a unique root of F. In this case, any point of I approximates the root of F in I.

2.3 Smale's α-theory

In this section, we recall Smale's α-theory, which is used to certify the solutions of square systems of analytic functions. Let $F : U \to \mathbb{C}^n$ be a square system of analytic functions

defined on open set $U \subset \mathbb{C}^n$. *Quadratic convergence* of $\{N_F^k(x)\}$ to a solution of F is defined as follows:

Definition 12. A point $x \in \mathbb{C}^n$ is called *an approximate solution* to F with *associated solution* x^* with $F(x^*) = 0$ if for every $k \in \mathbb{N}$,

$$\left\|N_F^k(x) - x^*\right\| \leq \left(\frac{1}{2}\right)^{2^k-1} \|x - x^*\|.$$

If $F'(x)$ is not invertible, then x is an approximate solution if and only if $F(x) = 0$.

α-theory provides a certificate for a point x to be an approximate solution to F using three values: $\alpha(F, x), \beta(F, x)$ and $\gamma(F, x)$. If $F'(x)$ is invertible, we define

$$\begin{aligned}
\alpha(F, x) &:= \beta(F, x)\gamma(F, x) \\
\beta(F, x) &:= \|x - N_F(x)\| = \|F'(x)^{-1}F(x)\| \\
\gamma(F, x) &:= \sup_{k\geq 2} \left\| \frac{F'(x)^{-1}F^{(k)}(x)}{k!} \right\|^{\frac{1}{k-1}}
\end{aligned} \tag{2.5}$$

where $F^{(k)}(x)$ in the definition of $\gamma(F, x)$ is a symmetric tensor whose components are the k-th partial derivatives of F, see [Lan83, §5]. The norm in $\beta(F, x)$ is the usual Euclidean norm and the norm in $\gamma(F, x)$ is the operator norm on $S^k\mathbb{C}^n$ (for details, see [HS12]). When F' is not invertible at x, we define $\alpha(F, x) = \beta(F, x) = \gamma(F, x) = \infty$, but we do not consider this case. The following theorem is the main theorem of α-theory:

Theorem 13. ([HS12, Theorem 2]) Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be a system of analytic functions, and let x be any point in $\mathbb{C}^n$. If

$$\alpha(F, x) < \frac{13 - 3\sqrt{17}}{4},$$

then x is an approximate solution for F. Moreover, $\|x - x^*\| \leq 2\beta(F, x)$ where x^* is the associated solution to x.

Moreover, with a stricter test, α-theory also provides a way to identify when other points approximate the same root of F. This is expressed in the following theorem:

Theorem 14 ([Blu+12, Theorem 4 and Remark 6, §8])**.** Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be a system of analytic functions, and let x be any point in $\mathbb{C}^n$. If

$$\alpha(F, x) < 0.03 \quad \text{and} \quad \|x - y\| < \frac{1}{20\gamma(F, x)},$$

then x and y are approximate solutions to the same root of F. Also, there is a unique root x^* of F in the ball centered at x with radius $\frac{1}{20\gamma(F,x)}$. Furthermore, if $\|x - \overline{x}\| > 4\beta(F, x)$, then x^* is not real.

Remark 15. The results of Theorems 13 and 14 apply when $\mathbb{C}$ is replaced by $\mathbb{R}$. In particular, if x is real and both F and F' are real-valued over $\mathbb{R}$, then, when the hypotheses of these theorems are satisfied, the corresponding root of F is real.

We observe that in many cases, β can be explicitly computed or bounded. For example, suppose there are oracles $\square F(x)$ and $\square F'(x)$ that return intervals or boxes containing $F(x)$ and $F'(x)$. Then, $\beta(F, x)$ can be estimated by bounding $\square F'(x)^{-1}\square F(x)$. In §2.5.2, we show that such oracles exist for D-finite functions. Therefore, throughout the remainder of this section, we focus on bounding the value of $\gamma(F, x)$.

In most applications of α-theory the key step is to compute (or bound) γ. In this section, we recall the construction in [SS00, §I-3] for the case where $F = P$ is a square polynomial system, i.e. $m = 0$ in Equation (2.1). These bounds are needed for the polynomial part for the general case of Equation (2.1).

For a polynomial $p = \sum_{|\nu| \le d} a_\nu x^\nu$, we recall that the Bombieri-Weyl norm is defined as

$$\|p\|^2 = \frac{1}{d!} \sum_{|\nu| \le d} \nu!(d - |\nu|)!|a_\nu|^2.$$

For a system of polynomials $P = (p_1, \ldots, p_n)$, we define

$$\|P\|^2 = \sum_{i=1}^{n} \|p_i\|^2.$$

Moreover, we let $d_i = \deg p_i$ be the degree of the i^{th} polynomial and $d = \max d_i$ be the maximum degree of the polynomials. For a point $x \in \mathbb{C}$, we denote $1 + \sum_{i=1}^{n} |x_i|^2$ by $\|(1, x)\|^2$, and we let $\Delta_P(x)$ be the diagonal matrix with entries

$$\Delta_P(x)_{ii} := \sqrt{d_i}\|(1, x)\|^{d_i - 1}.$$

With these definitions in hand, we may use them to bound γ for a polynomial system as follows:

Proposition 16 ([HS12, Proposition 5])**.** Let P be a square system of polynomials and suppose that $P'(x)$ is nonsingular at $x \in \mathbb{C}^n$. Define

$$\mu(P, x) := \max\left\{1, \|P\|\|P'(x)^{-1}\Delta_P(x)\|\right\}$$

where the norm in $\|P'(x)^{-1}\Delta_P(x)\|$ is the operator norm. Then,

$$\gamma(P, x) \leq \frac{\mu(P, x) d^{\frac{3}{2}}}{2\|(1, x)\|}.$$

2.4 Certifying solutions of polynomial systems

Based on two different methods, we certify regular roots of polynomial systems. For polynomial systems, the Krawczyk method and α-theory are implemented as a package `NumericalCertification` [Lee19] in `Macaulay2` [GS]. The readers who seek for a stand-alone software package can refer to `alphaCertified` [HS11]. We introduce how the package works and we apply this to certify approximations of regular roots obtained by a `Macaulay2` package `MonodromySolver` [Duf+].

2.4.1 Implementation : NumericalCertification

We introduce the Macaulay2 package NumericalCertification. As we suggested two methods in the previous section, the main functionality of our package is certifying regular roots of the square polynomial system using these. We have two main functions krawczykMethod and certifySolution which respectively uses the Krawczyk method and α-theory. In order to define an input polynomial systems we use the function polySystem in the package NumericalAlgebraicGeometry [Ley11]. The package supports the computation for $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ and Gaussian rational $\mathbb{Q}(i)$, and we display the examples on $\mathbb{C}$. As a running example we use the system from Example 3.

```
i1 : R = CC[x,y,z];
i2 : F = polySystem {x^3 - 3*x^2*y + 3*x*y^2 - y^3 - z^2,
                     z^3 - 3*z^2*x + 3*z*x^2 - x^3 - y^2,
                     y^3 - 3*y^2*z + 3*y*z^2 - z^3 - x^2};
```

We use an approximation to one of roots of F

$$(x, y, z) \approx (-.142332 - .358782i, .142332 - .358782i, .151879i)$$

obtained by NumericalAlgebraicGeometry.

2.4.1.1 krawczykMethod

In order to implement interval arithmetic in Macaulay2, we define a new class of data type Interval which can be defined by a function interval. Let us consider the above system and an interval vector consisting of intervals of radius .001 centered at (x, y, z).

```
i3 : (I1, I2, I3) = toSequence apply(o8, i ->
     interval(i-0.001-0.001*ii,i+0.001+0.001*ii));
```

Using a function `intervalOptionList`, we substitute intervals into corresponding variables.

```
i4 : o = intervalOptionList {("x" => "I1"), ("y" => "I2"),
("z" => "I3")};
i5 : krawczykMethod(F,o)
given interval contains a unique solution
o5 = true
```

2.4.1.2 certifySolution

Using a function `point` in `NumericalAlgebraicGeometry` we define an approximation for a root of the system.

```
i6 : p1 = point{{-.142332-.358782*ii,
                 .142332-.358782*ii, .151879*ii}};
i7 : p2 = point{{(-3.38813e-20-4.23516e-20*ii,
                 -3.38813e-20-4.23516e-20*ii,
                 -3.38813e-20-4.23516e-20*ii}};
          -- an approximation for a multiple root
i8 : P = {p1,p2};
```

In addition to `p1`, we define `p2` an approximation for the singular solution (at the origin) to observe how the package reacts in the case of failure. Then, it shows the following certification result:

```
i9 : certifySolution(F,P)
o9 = ({p1}, {(1.09874e-11, 7.44879e-14, 147.505)})
o9 : Sequence
```

It shows that `p1` is the only approximate solution of F. Also, we have the values of three parameters $(\alpha(F,x), \beta(F,x), \gamma(F,x)) = (1.09874 \cdot 10^{-11}, 7.44879 \cdot 10^{-14}, 147.505)$. Note

that the implementation does not prove the correctness of `p2` even though it is close to an actual root at the origin. The way to certify this approximation will be introduced in §3.

2.4.1.3 An application to `MonodromySolver`

In this section, for an application of `NumericalCertification`, we consider the numerical solver `MonodromySolver` [Duf+] in `Macaulay2` which implements the algorithms introduced in §1.4. The package `MonodromySolver` uses the functionality of the package `NumericalAlgebraicGeometry` [Ley11].

Note that the algorithms we use are driven partly by heuristics. As a post-processing step, we can certify the completeness and correctness of the solution set to a polynomial system computed with our main method. This is possible in the scenario when

- the parameteric system is square,
- all solutions are regular (the Jacobian of the system is invertible), and
- the solution count is known.

We can use Smale's α-theory to certify an approximation to a regular solution of a square system. Using a function `certifySolutions`, we determine whether the given solution is an approximate zero of the given polynomial system.

In the following example, all arithmetic and linear algebra operations are done over the field of Gaussian rationals $\mathbb{Q}(i)$. To use this certification method we first convert the coefficients of the system to Gaussian rationals, then perform certification numerically.

Example 17 (Nash equilibria)**.** Semi-mixed multihomogeneous systems arise when one is looking for all totally mixed Nash equilibria (TMNE) in game theory. A specialization of mixed volume using matrix permanents gives a concise formula for a root count for systems arising from TMNE problems [EV14]. We provide an overview of how such systems are constructed based on [EV14]. Suppose there are N players with m options each. For

player $i \in \{1, \ldots, N\}$ using option $j \in \{1, \ldots, m\}$ we have the equation $P_j^{(i)} = 0$, where

$$P_j^{(i)} = \sum_{\substack{k_1,\ldots,k_{i-1},\\ k_{i+1},\ldots,k_N}} a^{(i)}_{k_1,\ldots,k_{i-1},j,k_{i+1},\ldots,k_N} p_{k_1}^{(1)} p_{k_2}^{(2)} \cdots p_{k_{i-1}}^{(i-1)} p_{k_{i+1}}^{(i+1)} \cdots p_{k_N}^{(N)}. \tag{2.6}$$

The parameters $a^{(i)}_{k_1,k_2,\ldots,k_N}$ are the payoff rates for player i when players $1, \ldots, i-1, i+1, \ldots, N$ are using options $k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_N$, respectively. Here the unknowns are $p_{k_j}^{(i)}$, representing the probability that player i will use option $k_j \in \{1, \ldots, m\}$. There is one constraint on the probabilities for each player $i \in \{1, \ldots, N\}$, namely the condition that

$$p_1^{(i)} + p_2^{(i)} + \cdots + p_m^{(i)} = 1. \tag{2.7}$$

The system (2.6) consists of Nm equations in Nm unknowns. Condition (2.7) reduces the number of unknowns to $N(m-1)$. Lastly, we eliminate the $P_j^{(i)}$ by constructing

$$P_1^{(i)} = P_2^{(i)},\ P_1^{(i)} = P_3^{(i)},\ \ldots\ , P_1^{(i)} = P_m^{(i)}, \quad \text{for each } i \in \{1, \ldots, N\}.$$

The final system is a square system of $N(m-1)$ equations in $N(m-1)$ unknowns.

We chose the generic system of this form for $N = 3$ players with $m = 3$ options for each (see `paper-examples/example-Nash.m2` at [Duf+]). The result is a system of six equations in six unknowns and 81 parameters with 10 solutions. We also use `NumericalCertification` to demonstrate that these solutions can be certified.

2.5 Certifying solutions of systems of analytic functions

In this section, we extend our setting into systems with analytic functions. We take the two methods introduced earlier and apply them to a system (2.1) with the class of functions built from the coordinate functions and D-finite functions. We recall that a *D-finite function* g is a solution to a linear differential equation with polynomial coefficients $p_k(t) \in \mathbb{C}[t]$, i.e.

a differential equation of the following form:

$$p_r(t)g^{(r)}(t) + \cdots p_1(t)g'(t) + p_0(t)g(t) = 0. \tag{2.8}$$

If $p_r(0)$ does not vanish, then there is a unique function $g(t)$ which satisfies both Equation (2.8) and specified initial conditions $g(0) = c_0$, $g'(0) = c_1$, ..., and $g^{(r-1)}(0) = c_{r-1}$. We call the corresponding class of functions *polynomial-D-finite functions*.

2.5.1 α-theory on an analytic system

We apply the results from §2.3 to systems of the form of Equation (2.1). In particular, we call P the part of F consisting of polynomial equations. We begin by observing that the results in [HL17, Theorem 2.3] can be directly generalized to the setting of analytic functions. In particular, we let

$$\Delta_F = \begin{bmatrix} \Delta_P(x)\|P\| & & \\ & & \\ & & I_m \end{bmatrix}$$

be an $(n+m) \times (n+m)$ diagonal matrix. When F' is invertible at $x \in \mathbb{C}^{n+m}$, we define

$$\mu(F, x) := \max\left\{1, \left\|F'(x)^{-1}\Delta_F\right\|\right\}.$$

By the proof of [HL17, Theorem 2.3], we conclude that

$$\gamma(F, x) \leq \mu(F, x) \sup_{k\geq 2} \left(\left(\frac{d^{\frac{3}{2}}}{2\|(1, x)\|} \right)^{2(k-1)} + \sum_{i=1}^{m} \left| \frac{g_i^{(k)}(x_i)}{k!} \right|^2 \right)^{\frac{1}{2(k-1)}}.$$

By the concavity of the of the $2(k-1)^{\text{th}}$ root, it follows that

$$\gamma(F, x) \leq \mu(F, x) \left(\frac{d^{\frac{3}{2}}}{2\|(1, x)\|} + \sup_{k\geq 2} \sum_{i=1}^{m} \left| \frac{g_i^{(k)}(x_i)}{k!} \right|^{\frac{1}{k-1}} \right). \tag{2.9}$$

Therefore, we observe that, in order to get a bound on γ, it is enough to bound $\left|\frac{g_i^{(k)}(t)}{k!}\right|^{\frac{1}{k-1}}$ independently of k for each ingredient g_i. In [HL17], Hauenstein and Levandovskyy find a bound on these quantities using a recurrence relation from the defining linear differential equation with constant coefficients. We achieve such a bound via the Cauchy integral theorem.

Lemma 18. Suppose that the following two oracles exist:

1. Given a univariate analytic function g and a point $x \in \mathbb{C}$ in the domain of g, there is an oracle which returns a positive value $R > 0$ so that the radius of convergence of a power series for g centered at x is at least R.

2. Given a univariate analytic function g, a point $x \in \mathbb{C}$ in the domain of g, and a radius r, there is an oracle which returns M which is an upper bound on the value of $|g|$ on the closed disk $\overline{D}(x, r)$.

Then, for $k \geq 2$,

$$\left|\frac{g^{(k)}(t)}{k!}\right|^{\frac{1}{k-1}} \leq \frac{1}{r} \max\left\{1, \frac{M}{r}\right\}.$$

Proof. Using Cauchy's integral theorem, we have that

$$\frac{|g^{(k)}(x)|}{k!} = \left|\int_0^1 \frac{g(x + re^{2\pi it})}{(re^{2\pi it})^k} dt\right| \leq \frac{M}{r^k}.$$

Therefore,

$$\left|\frac{g^{(k)}(x)}{k!}\right|^{\frac{1}{k-1}} \leq \frac{1}{r}\left(\frac{M}{r}\right)^{\frac{1}{k-1}}.$$

Since $k \geq 2$, the $(k-1)^{\text{th}}$ root of $\frac{M}{r}$ is bounded as in the statement of the lemma. □

From this bound, which is independent of k, we can now derive a bound on $\gamma(F, x)$. By substituting this formula into Inequality (2.9), we have a bound on $\gamma(F, x)$. We collect this result in the following theorem:

Theorem 19. Let $U \subset \mathbb{C}^{n+m}$ and consider a system $F : U \to \mathbb{C}^{n+m}$ as in Equation (2.1) and let $x \in \mathbb{C}^{n+m}$. Moreover, assume that there exist oracles as in the statement of Lemma 18. For each g_i, let R_i be a positive lower bound on the radius of convergence for g_i at x_i (given by the first oracle in Lemma 18). For each i, fix $0 < r_i < R_i$ to be a positive value strictly less than the radius of convergence. Then, using the second oracle in Lemma 18, let M_i be an upper bound on $|g_i|$ on the closed disk $\overline{D}(x_i, r_i)$. For each i, let

$$C_i = \frac{1}{r_i} \max\left\{1, \frac{M_i}{r_i}\right\}.$$

Then,

$$\gamma(F, x) \leq \mu(F, x) \left(\frac{d^{\frac{3}{2}}}{2\|(1, x)\|} + \sum_{i=1}^{m} C_i \right).$$

Remark 20. We remark that the choice of r_i is critically important in this computation. When r_i is small, $\frac{1}{r_i}$ becomes large, and when r_i is quite large, the disk $\overline{D}(x, r_i)$ approaches a singularity of g_i, so M_i is quite large. Therefore, different choices of r_i can affect the value of C_i drastically. We provide experimental data illustrating this issue in §2.5.3.

We observe that we may apply the same approach as in Theorem 19 to both g' and g'' to achieve potentially tighter bounds on $\gamma(F, x)$. We make this explicit in the following corollary:

Corollary 21. Suppose that the conditions of Theorem 19 hold and, in addition, the oracles in the statement of Lemma 18 exist for both g' and g''. Let M_i' and M_i'' be upper bounds on $|g_i'|$ and $|g_i''|$, respectively, given by the oracle from Lemma 18 on $\overline{D}(x_i, r_i)$. Then, the C_i in Theorem 19 can be replaced by

$$C_i = \frac{1}{r_i} \max\left\{1, \min\left\{\frac{M_i}{r_i}, \frac{M_i'}{2}, \frac{M_i'' r_i}{2}\right\}\right\}. \tag{2.10}$$

Proof. We illustrate the key step in the computation for M_i'; the other cases are similar or

appear in Theorem 19. We observe that since $k \geq 2$,

$$\left|\frac{g_i^{(k)}(x)}{k!}\right|^{\frac{1}{k-1}} \leq \left|\frac{(g_i')^{(k-1)}(x)}{2(k-1)!}\right|^{\frac{1}{k-1}} \leq \frac{1}{r_i}\left(\frac{M_i}{2r_i}\right)^{\frac{1}{k-1}},$$

where the second inequality follows from applying the inequality of Lemma 18 to g'. By considering the possible magnitudes of $\frac{M_i}{2r_i}$, the desired result follows. □

Based on the discussion above we outline an algorithm to certify a root of the system F.

AlphaTest$(F, x, r_i, M_i, M_i', M_i'')$**:**

Input: A differentiable system of functions $F : U \to \mathbb{C}^{n+m}$ for an open set $U \subset \mathbb{C}^{n+m}$, a point $x \in \mathbb{C}^{n+m}$, a positive value r_i such that $0 < r_i < R_i$ for each i, and upper bounds M_i, M_i', M_i'' on $|g_i|, |g_i'|, |g_i''|$ on the closed disk $\overline{D}(x_i, r_i)$ for each i.

Output: The boolean value of a condition that implies "x is an approximate solution of F".

Compute constants $\beta(F, x), \mu(F, x)$ and C_i in Corollary 21.

return $\beta(F,x)\mu(F,x)\left(\frac{d^{\frac{3}{2}}}{2\|(1,x)\|} + \sum_{i=1}^{m} C_i\right) < \frac{13-3\sqrt{17}}{4}$,

---

In the next section, we show that the oracles required by Lemma 18 exist for D-finite functions.

Remark 22. We observe that the results in this section apply when $\mathbb{C}$ is replaced by $\mathbb{R}$. In particular, real roots are certified using the standard techniques of α-theory for real roots. The derived bounds on γ, however, use the complex values of the radius of convergence and maximum of the function, not merely the real part.

2.5.2 The case of D-finite functions

In this section, we show that the oracles needed in §§2.2 and 2.3 exist for D-finite functions. These oracles fall into two classes: evaluating a D-finite function or finding the radius of convergence of a D-finite function. We point out that the oracles can be obtained from known software implementations.

2.5.2.1 *Evaluating D-finite functions*

The analytic continuation algorithm of Chudnovsky and Chudnovsky, first presented in [CC90] and further developed in [Hoe99], provides an algorithm to approximate the value of a D-finite function. In particular, the package `ore_algebra.analytic` [Mez16] in `SageMath` [The18] uses this technique and provides functions which compute an interval containing the image of a D-finite function over a point or interval.

The output of this algorithm can be used to calculate intervals or boxes containing F and F' when evaluated at points or over intervals (we note that the derivative of a D-finite function is also D-finite).

Remark 23. In the real case, an alternate approximation method using Chebyshev polynomials is presented in [BJM17]. These methods return Chebyshev polynomials such that, on an interval I, the point-wise difference between the polynomial and the prescribed D-finite function is within a specified error. By applying interval arithmetic on this polynomial, a D-finite function can be evaluated on an interval. An implementation of this approximation is available in `Maple` [Map18] and experimental source code is referenced in [BJM17].

2.5.2.2 *The radius of convergence for D-finite functions*

Mezzarobba and Salvy present an algorithm to compute the majorant series for D-finite function in [MS10]. In this case, the radius of convergence for the majorant series is a lower bound on the radius of convergence for the corresponding D-finite function. The majorant series provided in [MS10] has a particularly simple presentation, where the radius of convergence can be identified by the vanishing of a linear term of in denominator, see [MS10, Equation (18)]. The `Maple` package `numGfun` [Mez10] and the `SageMath` [The18] package `ore_algebra.analytic` provide algorithms for computing this majorant series. For extensions and details of the majorant series approach, see [MP98] and [Hoe03].

2.5.3 Experiments

In this section, as a proof of concept, we provide some computational and experimental results for our certification methods for D-finite functions, as described in §2.5.2. Our implementations are in `SageMath` [The18]. We use the `ore_algebra.analytic` package from [Mez16] for evaluation of D-finite functions (function `numerical_solution`) and for estimating the radius of convergence for the majorant series of D-finite functions (function `leading_coefficient`). The code and all examples in this section are available at

`https://github.com/klee669/DfiniteComputationResults`

2.5.3.1 Comparison between α-theory and the Krawczyk method.

The *error function* $\operatorname{erf}(t)$ is a basic example of a D-finite function which satisfies the following differential equation and initial conditions:

$$\operatorname{erf}''(t) + 2t\operatorname{erf}'(t) = 0, \quad \operatorname{erf}(0) = 0, \quad \operatorname{erf}'(0) = \frac{2}{\sqrt{\pi}}.$$

We note that the error function has no singularities in $\mathbb{C}$. We consider the following square system of equations along with the corresponding square function F.

$$\left\{ \begin{array}{r} t_1^2 + t_2^2 = 4 \\ 2\operatorname{erf}(t_1)\operatorname{erf}(t_2) = 1 \end{array} \right\} \text{ with } F(t_1,t_2,t_3,t_4) = \begin{bmatrix} t_1^2 + t_2^2 - 4 \\ t_3 t_4 - \frac{1}{2} \\ t_3 - \operatorname{erf}(t_1) \\ t_4 - \operatorname{erf}(t_2) \end{bmatrix}. \tag{2.11}$$

Using `Mathematica` [Wol], we find the following potential solution to this system of equations:

$$t = (t_1, t_2, t_3, t_4) = (.480322, 1.94147, .503058, .993961). \tag{2.12}$$

Using both α-theory and the Krawczyk method, we certify that this point approximates a solution to the system of equations in Equation (2.11). In order to study the accuracy required for the α-theory-based and Krawczyk method-based tests, we round the coordinates of the point in Equation (2.12) to d decimal places and vary d in our experiments appearing in Table 2.1. For Krawczyk method-based tests, we also to specify a region by choosing the box whose side length is 2×10^{-d} centered at the rounded approximation. Moreover, we choose $F'(m(I))^{-1}$ as the invertible matrix Y in Equation (2.4).

Table 2.1: Comparison between the precision required for the Krawczyk-based and $\boldsymbol{\alpha}$-theory-based methods.

decimal places	Krawczyk method	α-theory
0	fail	fail
1	pass	fail
2	pass	fail
3	pass	pass

For the Krawczyk method, a pass indicates that the generalization of the Newton operator is contractive within the given region using the test described in §2.2. On the other hand, for the α-theory-based test from §2.3, a pass indicates that the approximation is certified to be an approximate solution. Throughout this example, we use $r = 0.4$ for the α-theory-based test as that gives (nearly) the best value for r_i, cf. Remark 20.

We observe that Equation (2.11) is an example of a system which could not be effectively studied using the previous α-theory techniques. We also note that the Krawczyk method succeeds with less precision than the α-theory-based test. This behavior is not surprising as the Krawczyk method has a weaker convergence result and uses less pessimistic estimates in its computation.

2.5.4 The radius for the α-theory-based test.

In this section, we provide some experimental data illustrating the care that must be taken in choosing the radius from §2.3, see Remark 20. We consider a Bessel function (of order ν) $y(t) = C_1 Y_\nu(t) + C_2 J_\nu(t)$. This function is a D-finite function satisfying the following differential equation:

$$t^2 y''(t) + t y'(t) + (t^2 - \nu^2) y(t) = 0.$$

We consider the case where $\nu = 9$. In this case, the Bessel function has a regular singularity at $t = 0$, its derivative has singularities at $t = 0, \pm 9$, and the second derivative has singularities at $t \approx 0, \pm 8.2923, \pm 9, \pm 9.7076$. Consider the following system of equations and corresponding system F involving a Bessel function and an error function.

$$\left\{ \begin{array}{r} t_1^2 + t_2^2 = 61 \\ 2\,\mathrm{erf}\left(\frac{1}{2}\left(Y_9(t_2) + J_9(t_2)\right) + t_1\right)\left(Y_9(t_2) + J_9(t_2)\right) = 11 \end{array} \right\} \text{ with}$$

$$F(t_1, t_2, t_3, t_4, t_5) = \begin{bmatrix} t_1^2 + t_2^2 - 61 \\ 2t_4 t_5 - 11 \\ t_3 - \frac{1}{2} t_5 - t_1 \\ t_4 - \mathrm{erf}(t_3) \\ t_5 - \left(Y_9(t_2) + J_9(t_2)\right) \end{bmatrix}.$$

Using `Mathematica` [Wol], we find the following potential solution to this system of equations:

$$t = (t_1, t_2, t_3, t_4, t_5) = (6.27899, 4.64481, -.38382, -.41274, -13.32563).$$

We apply the α-theory-based method of §2.3 in attempt to certify this solution while varying radii (up to $3 \cdot 10^{-2} R$ which is close to the maximal computable radius) using

the experimentally found lower bound for the radius of convergence, $R = 8.2923$. We summarize our results in Table 2.2.

Table 2.2: $\gamma(F,t)$ and $\alpha(F,t)$ values depending on radii.

radius	$\gamma(F,t)$	$\alpha(F,t)$	passes α-test?
$10^{-6}R$	$6.9909 \cdot 10^7$	1.3078	no
$10^{-5}R$	$6.9909 \cdot 10^6$	.1308	yes
$10^{-4}R$	$6.9912 \cdot 10^5$	.0131	yes
$10^{-3}R$	$2.2485 \cdot 10^5$	.0042	yes
$10^{-2}R$	$1.9722 \cdot 10^6$	.0369	yes
$3 \cdot 10^{-2}R$	$2.4525 \cdot 10^7$	.4588	no

We observe that, as expected from Remark 20, a radius which is either too small or too large (when compared to the distance to the singularity) can result in a need for increased precision in the α-theory-based test.

2.5.4.1 Comparing α-theory-based tests on polynomial-exponential systems

In this section, we compare the bounds on γ that we derive to those from the polynomial-exponential systems in [HS12]. In particular, we consider the following example in the class of polynomial-exponential systems (which are a special case of polynomial-D-finite systems):

$$\left\{e^{4t} = .0183\right\} \text{ with } F(t_1,t_2) = \begin{bmatrix} t_2 - .0183 \\ t_2 - e^{4t_1} \end{bmatrix}.$$

For the approximate solution $(t_1,t_2) = (-1, .018316)$, we compare the bounds on $\gamma(F,t)$ for the method presented in this thesis to the γ from [HS12], as computed by the software `alphaCertified` [HS11]. We separate out the three bounds on γ from Corollary 21. Figure 2.1 compares the results from `alphaCertified` and our method. There, we see that both theoretically and in our implementation, the computed γ-value may be less than that in [HS12], as computed by `alphaCertified`. We note that, in Figure 2.1, the implementation bounds differ from the theoretical bounds because the package `ore_algebra.analytic` returns inexact outputs when it evaluates functions over an interval.

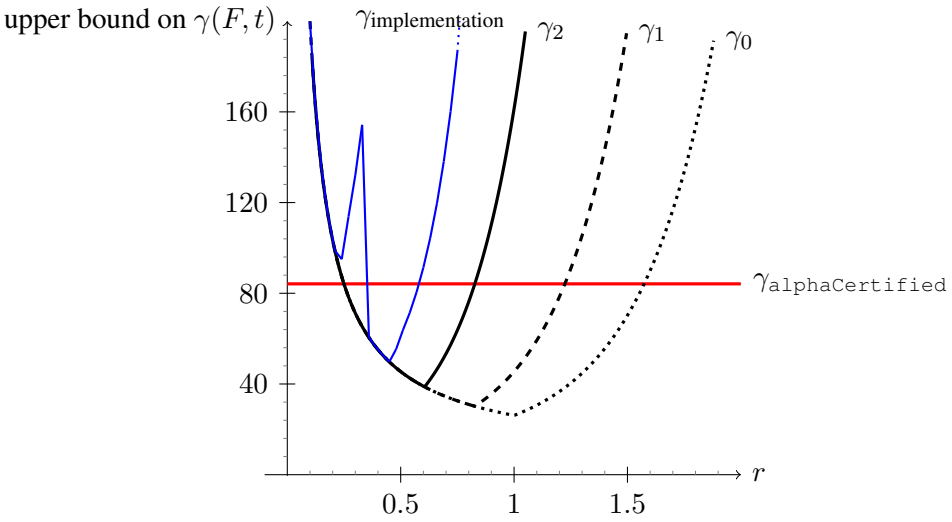

Figure 2.1: Comparison of computed γ values in this paper to those from the software `alphaCertified`. $\gamma_0, \gamma_1, \gamma_2$ indicate bounds computed through $\frac{M_i}{r_i}$, $\frac{M_i'}{2}$, $\frac{M_i'' r_i}{2}$ in (2.10) respectively. $\gamma_{\text{implementation}}$ indicates bounds computed by the implementation. $\gamma_0, \gamma_1, \gamma_2$ and $\gamma_{\text{implementation}}$ have lower values of bounds than `alphaCertified` for some choices of r.

2.5.4.2 Application to an optimization problem

We also use our implementation to solve an optimization problem involving the perimeters of ellipses. Suppose that E_1, E_2 are ellipses with major axes of lengths 1 and 2, respectively, whose perimeters sum to 17. Suppose that we want to maximize

$$e_1 A_1 + e_2 A_2$$

where e_i is the eccentricity and A_i is the area of E_i. Since the area of an ellipse is the product of π and the lengths of its axes, if we let b_i be the length of the minor axis of E_i,

this maximization problem is equivalent to the following problem:

$$\begin{aligned} \text{Maximize} \quad & e_1 b_1 + 2 e_2 b_2 \\ \text{subject to} \quad & e_1^2 + b_1^2 = 1 \\ & 4 e_2^2 + b_2^2 = 4 \\ & 4E(e_1) + 8E(e_2) = 17 \end{aligned}$$

where $E(t) = \int_0^1 \frac{\sqrt{1-t^2x^2}}{\sqrt{1-x^2}} dx$ is the complete elliptic integral of the second kind, which satisfies the differential equation

$$(t - t^3)E''(t) + (1 - t^2)E'(t) + tE(t) = 0.$$

Liouville [Lio40] showed that $E(t)$ is not algebraic. We can rewrite this maximization problem as a square system of equations by setting up a Lagrange multiplier system. Since derivatives of D-finite functions are still D-finite functions [Hoe99], the square system can be certified by the α theory- and the Krawczyk method-based approaches. In our experiments, we use the approximate solution

$$\begin{aligned} & (b_1, b_2) = (.8337853, 1.5601133), \\ & (e_1, e_2) = (.5520888, .6257089) \quad \text{and} \\ & (\lambda_1, \lambda_2, \lambda_3) = (-.3310737, -.4010663, .0590727) \end{aligned}$$

With the choice of the radius $r = 0.01$ and an approximate solution with 7 digits of precision, the α-theory-based test certifies the approximate solution. On the other hand, Krawczyk method-based test requires much less precision, in fact, only 2 digits of precision and a box of side length 2×10^{-2} are enough to certify this solution.

CHAPTER 3

CERTIFYING MULTIPLE SOLUTIONS OF SYSTEMS OF EQUATIONS

In this section, we consider the *local separation bound* of isolated multiple solutions of square systems of equations. In other words, for a given multiple root x^* of a square analytic system F, we find the minimum distance between x^* and other roots of F. The local separation bound is important for providing an upper bound on the number of steps that subdivision-based algorithms perform in order to isolate x^* from other roots of F. It also provides a criterion of the quadratic convergence of Newton's method. The separation bound for roots of multiplicity 2 was studied in [DS01] and roots with Jacobian of corank 1 was done in [Hao+20]. In this section, we focus on the local separation bound for a simple multiple root, i.e. an isolated multiple root of a system satisfying that the deflation algorithm applied on the system and the root terminates after only one iteration. Also, we use the separation bound for certifying an approximation of an isolated multiple root of a system.

3.1 Preliminaries

In this section, we introduce the concepts required to define the multiple roots of our interest. First, we start with the general notion of the local dual space in order to describe the multiplicity structure. From this concept, we define the multiplicity of a zero for a polynomial system and find its lower bound when the zero is isolated. Secondly, a deflation method is suggested which plays an important role to define an explicit family of multiple roots that we want to certify their approximations. Note that these are defined and applied over polynomial systems. Therefore, in this section, we restrict arguments to the system consisting of polynomials. We extend the setting to any square analytic system in §3.2.

3.1.1 Local dual space and multiplicity

The local dual space is a tool to analyze the multiplicity of a zero for a system of equations. Let $\mathbf{d}_{x^*}^{\boldsymbol{\alpha}} : \mathbb{C}[\mathbf{x}] = \mathbb{C}[x_1, \ldots, x_n] \to \mathbb{C}$ denote the differential functional defined by

$$\mathbf{d}_{x^*}^{\boldsymbol{\alpha}}(g) = \frac{1}{\alpha_1! \cdots \alpha_n!} \cdot \frac{\partial^{|\boldsymbol{\alpha}|} g}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}(x^*), \quad \forall g \in \mathbb{C}[\mathbf{x}],$$

where $x^* \in \mathbb{C}^n$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n] \in \mathbb{N}^n$. Let I_F denote the ideal generated by $F = \{f_1, \ldots, f_n\}$, where $f_i \in \mathbb{C}[\mathbf{x}]$. The local dual space of I_F at an isolated zero x^* is a subspace of $\mathfrak{D}_{x^*} = \mathrm{span}_{\mathbb{C}}\{\mathbf{d}_{x^*}^{\alpha}\}$

$$\mathcal{D}_{F,x^*} = \{\Lambda \in \mathfrak{D}_{x^*} \mid \Lambda(g) = 0, \ \forall g \in I_F\}.$$

Let $\mathcal{D}_{F,x^*}^{(k)}$ denote the subspace of $\mathcal{D}_{F,x^*}$ with differential functionals of order bounded by k, we define

1. breadth $\kappa = \dim\left(\mathcal{D}_{F,x^*}^{(1)}\right) - \dim\left(\mathcal{D}_{F,x^*}^{(0)}\right) \equiv \dim \ker F'(x^*)$,
2. depth $\rho = \min\left\{k \mid \dim\left(\mathcal{D}_{F,x^*}^{(k+1)}\right) = \dim\left(\mathcal{D}_{F,x^*}^{(k)}\right)\right\}$,
3. multiplicity $\mu = \dim\left(\mathcal{D}_{F,x^*}^{(\rho)}\right)$.

If x^* is an isolated multiple zero of F, then $1 \leq \kappa \leq n$ and $\rho < \mu < \infty$. Define the function

$$H(k) = \begin{cases} \dim\left(\mathcal{D}_{F,x^*}^{(0)}\right) \equiv 1 & \text{if } k = 0 \\ \dim\left(\mathcal{D}_{F,x^*}^{(k)}\right) - \dim\left(\mathcal{D}_{F,x^*}^{(k-1)}\right) & \text{otherwise.} \end{cases}$$

We call this function as the *Hilbert function* at x^*. The properties of the Hilbert function are summarized by the following lemma:

Lemma 24 (c.f. [Sta78; DZ05b])**.** Let $H(k)$ be the Hilbert function at x^*. Then, the following holds:

1. $H(k) = 0$ for sufficiently large k if and only if x^* is an isolated zero.

2. $H(k) \leq \binom{\kappa+k-1}{\kappa-1}$ for all k.

3. If $H(k) = 0$ for some k then $H(l) = 0$ for any $l \geq k$.

4. If $H(k) = 1$ for some $k > 0$, then $H(l) \leq 1$ for any $l \geq k$.

5. If $H(k) \leq k$ for some k, then a sequence $\{H(l)\}_{l=k}^{\infty}$ is non-increasing.

The multiplicity μ defined above is sometimes called as "arithmetical multiplicity" [MMM95]. It is closely related to the concept of the multiplicity from the commutative algebra point of view. With the same setting as above, we define "intersection multiplicity" of x^* as $\dim_{\mathbb{C}} (\mathbb{C}[\![\mathbf{x}]\!]/\langle f_1, \ldots, f_n \rangle)$ where $\mathbb{C}[\![\mathbf{x}]\!]$ is the ring of convergent power series at x^*. Then, the following theorem establishes the relationship between the arithmetical mulitplicity and the intersection multiplicity.

Theorem 25. [DZ05b, Theorem 1] Let I_F be an ideal in $\mathbb{C}[\mathbf{x}]$ generated by $F = \{f_1, \ldots, f_n\}$ such that F has an isolated zero x^*, then the intersection multiplicity of x^* equals to the arithmetical multiplicity of x^*.

As we only focus on isolated multiple zeros, we just call the two concepts of multiplicities as "multiplicity" without distinguishing them.

The minimum value of the multiplicity μ of a zero is obtained from its breadth. Suppose that we have the breadth κ, i.e. $\dim \ker F'(x^*) = \kappa$. For a local ring $(\mathfrak{A}, \mathfrak{m})$ and an $\mathfrak{A}$-module M, we denote the multiplicity of M with respect to $\mathfrak{m}$ by $\mu(M)$. We remind the following fact from commutative algebra.

Proposition 26. [Bou06, VIII, §7.4, Proposition 7] Suppose that $(\mathfrak{A}, \mathfrak{m})$ is a local ring. Let $s \geq 1$ be integer satisfying that for $1 \leq i \leq s$, there are an integer $\mathfrak{d}_i > 0$ and an element x_i in $\mathfrak{m}^{\mathfrak{d}_i}$ with its class ξ_i in $\mathfrak{m}^{\mathfrak{d}_i}/\mathfrak{m}^{\mathfrak{d}_i+1}$. Suppose that $\{x_1, \ldots, x_s\}$ is a regular sequence for $\mathfrak{A}$. If we denote $\mathbf{x}$ be an ideal of $\mathfrak{A}$ generated by $\{x_1, \ldots, x_s\}$, then $\mu(\mathfrak{A}/\mathbf{x}) \geq \mathfrak{d}_1 \cdots \mathfrak{d}_s \cdot \mu(\mathfrak{A})$.

We now apply this proposition to our setting. The next theorem provides a lower bound for the multiplicity of an isolated multiple root for a square polynomial system.

Theorem 27. Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be any square polynomial system and x^* be an isolated multiple root of F with $\dim \ker F'(x^*) = \kappa < n$. Then, the multiplicity of x^* is at least 2^κ.

Proof. Suppose that x^* has the multiplicity less than 2^κ. Without loss of generality, we may assume that x^* is the origin. First consider the local ring $\mathfrak{A} = \mathbb{C}[x_1, \ldots, x_n]_{\mathfrak{m}}$ where $\mathfrak{m} = \langle x_1, \ldots, x_n \rangle$. Let $\mathbf{f} = \langle \tilde{f}_1, \ldots, \tilde{f}_n \rangle$ be an ideal generated by a regular sequence $\{\tilde{f}_1, \ldots, \tilde{f}_n\} \subset \langle f_1, \ldots, f_n \rangle$ for $\mathfrak{A}$. Such a regular sequence exists because x is an isolated root. Since the multiplicity of x is less than 2^κ, we know that $\mu(\mathfrak{A}/\mathbf{f}) < 2^\kappa$. Therefore, $\{\tilde{f}_1, \ldots, \tilde{f}_n\}$ has at least $n - \kappa + 1$ linear order elements by Proposition 26. When we observe the basis elements of $\mathcal{D}_{F,x^*}$, there are at most $\kappa - 1$ basis elements in $\mathcal{D}^{(1)}_{F,x^*}$. This is a contradiction because we have $\dim \ker F'(x) = \kappa$. □

Remark 28.

1. The condition that the system is square is necessary. Consider a polynomial system
$$F(x, y) = \begin{bmatrix} x^2 \\ xy \\ y^2 \end{bmatrix}.$$
Then, F has an isolated multiple root at the origin with $\kappa = 2$ and $\mu = 3 < 2^\kappa$.

2. The multiplicity of an isolated multiple root can be arbitrarily large. For example, a system
$$F(x, y, z) = \begin{bmatrix} x^n - yz \\ y^n - xz \\ z^n - xy \end{bmatrix}, \quad n \geq 3$$
has an isolated multiple root at the origin and its multiplicity is $2 + 3n$.

The lower bound of the multiplicity will make an important contribution in our main results. When a system is perturbed, a multiple root becomes an cluster of zeros of multiplicity many points. We derive a separation bound for the cluster in §3.4. When we obtain a ball containing the cluster of zeros, the above theorem suggests a lower bound of the

number of roots inside the ball.

3.1.2 Deflation method

A *deflation* is a class of effective methods to reinstate the quadratic convergence of Newton's iteration in the case of isolated singular zeros of square polynomial system. The basic idea for a deflation method is to introduce extra equations from original singularities for generating augmented systems with reduced singularities (e.g. multiplicity). In particular, [LVZ06] proposes an effective deflation method, which can be described as follows.

Suppose we are given an isolated singular zero $x^* \in \mathbb{C}^n$ of a polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$, satisfying

1. $F(x^*) = 0$,
2. $\dim \ker F'(x^*) = \kappa > 0$,

where $F'(x^*)$ is the Jacobian matrix of F at x^*. The goal is constructing an augmented system having a root with smaller multiplicity than that of x^* by introducing new equations obtained from F'. Let $B \in \mathbb{C}^{n\times(n-\kappa+1)}$ be a random matrix and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{n-\kappa+1})^\top$ be any vector. Then, $B \cdot \boldsymbol{\lambda}$ represents a parametrization of an $n - \kappa + 1$-dimensional space in $\mathbb{C}^n$. As the dimension for the kernel of $F'(x^*)$ is κ, we have $\ker F'(x^*) \cdot B \cdot \boldsymbol{\lambda}$ which is 1-dimensional. In order to achieve a zero-dimensional solution space, one more linear equation is introduced. For a random vector $\mathbf{b} \in \mathbb{C}^{n-\kappa+1}$, we consider a generic linear equation $\mathbf{b}^\top\boldsymbol{\lambda} - 1$. Then, with probability one (generic pairs of B and $\mathbf{b}$ in $\mathbb{C}^{n\times(n-\kappa+1)} \times \mathbb{C}^{n-\kappa+1}$), there exists a unique vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{n-\kappa+1})^\top$ such that $(x^*, \boldsymbol{\lambda})$ is an isolated zero of an augmented system

$$G = \begin{bmatrix} F \\ F' \cdot B \cdot \boldsymbol{\lambda} \\ \mathbf{b}^\top \cdot \boldsymbol{\lambda} - 1 \end{bmatrix}.$$

If $(x^*, \boldsymbol{\lambda})$ remains a singular zero of G, the deflation process is repeated for G and $(x^*, \boldsymbol{\lambda})$.

In fact, the extra equations guarantee that $B \cdot \boldsymbol{\lambda}$ is a random nonzero sample from $\ker F'(x^*)$. In order to see this more clearly, we propose an equivalent deflation. Let $V = (V_1, V_2) \in \mathbb{C}^{n \times n}$, satisfying

1. $V_1 \in \mathbb{C}^{n \times \kappa}$, $\operatorname{im} V_1 = \ker F'(x^*)$,

2. $V_2 \in \mathbb{C}^{n \times (n-\kappa)}$, $\operatorname{im} V_2 = \{\ker F'(x^*)\}^{\perp}$,

and $\boldsymbol{\lambda}_1 = (\lambda_1, \ldots, \lambda_\kappa)^\top \in \mathbb{C}^\kappa$ be a random vector. Then with probability one (exceptional $\boldsymbol{\lambda}_1 = 0$), there exists a unique vector $\boldsymbol{\lambda}_2 = 0$ such that $(x^*, 0)$ is an isolated zero of an augmented system

$$G(x, \boldsymbol{\lambda}_2) = \begin{bmatrix} F \\ F' \cdot V \cdot \boldsymbol{\lambda} \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)^\top$. The Jacobian matrix of G at $(x^*, 0)$ is calculated

$$G'(x^*, 0) = \begin{bmatrix} F'(x^*) & \mathbf{0} \\ F''(x^*) \cdot V_1 \cdot \boldsymbol{\lambda}_1 & F'(x^*) \cdot V_2 \end{bmatrix} \in \mathbb{C}^{2n \times (2n-\kappa)},$$

where $F''(x^*)$ is the $n \times n \times n$ tensor consisting of all second order derivatives of F at x^*. If $G'(x^*, 0)$ remains singular, the deflation process is repeated for G and $(x^*, 0)$.

Leykin, Verschelde and Zhao proved that the number of deflation steps is strictly less than μ [LVZ06, Theorem 3.1]. Dayton and Zeng further proved that it is less than the depth ρ which is a tighter bound [DZ05a, Theorem 3]. Li and Zhi proved that the worst case bound is always true when $\kappa = 1$ [LZ12, Theorem 3.8]. However, by observing from the testing benchmark list in [DZ05a], the deflation process for many kinds of the systems with a multiple zero terminates by only one step when $\kappa > 1$. These isolated singular zeros are of our particular interest.

3.2 Simple multiple roots

In this section, we define singular zeros that we focus on. For a square polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$, we focus on its isolated multiple zero x^* satisfying the property that the deflation process applied on F and x^* terminates after only one iteration.

In order to deal with this 'one step deflation sufficient' zero x^* of F in a more tangible form, we find a characterization of it. T'his characterization should present a linear operator which can be used to prove the statements in the next section. We show that such a characterization exists.

Theorem 29. Let a point $x^* \in \mathbb{C}^n$ be an isolated multiple zero of a polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$. Define the square matrix

$$B = \begin{bmatrix} F''(x^*)(v_1, v_1) & \cdots & F''(x^*)(v_\kappa, v_\kappa) & F'(x^*) \cdot V_2 \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

Then, B is nonsingular for almost all choices of orthonormal bases $\{v_1, \ldots, v_\kappa, V_2\}$ with $\operatorname{im}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\operatorname{im} V_2 = \{\ker F'(x^*)\}^\perp$ if and only if the deflation process terminates after one step for almost all choices of bases $\operatorname{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \operatorname{im} V_1 = \ker F'(x^*)$ and $\boldsymbol{\lambda}_1 = (\lambda_1, \ldots, \lambda_\kappa)^\top \in \mathbb{C}^\kappa$, i.e.

$$G'(x^*, 0) = \begin{bmatrix} F'(x^*) & \mathbf{0} \\ F''(x^*) \cdot V_1 \cdot \boldsymbol{\lambda}_1 & F'(x^*) \cdot V_2 \end{bmatrix} \in \mathbb{C}^{2n \times (2n-\kappa)},$$

is of full rank.

Proof. We start with proving the "only if" direction. First, we prove that $G'(x^*, 0)$ is of full rank for almost all choices of bases $\operatorname{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\boldsymbol{\lambda}_1 = (\lambda_1, \ldots, \lambda_\kappa)^\top \in \mathbb{C}^\kappa$ is equivalent to the matrix

$$A = \begin{bmatrix} \sum_{i=1}^{\kappa} \lambda_i F''(x^*)(v_1, v_i) & \cdots & \sum_{i=1}^{\kappa} \lambda_i F''(x^*)(v_\kappa, v_i) & F'(x^*) \cdot V_2 \end{bmatrix}$$

is of full rank for almost all choices of bases $\mathrm{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\boldsymbol{\lambda}_1 = (\lambda_1, \ldots, \lambda_\kappa)^\top \in \mathbb{C}^\kappa$.

Let $w = (w_1, w_2)^\top$ where $w_1 \in \mathbb{C}^\kappa$, $w_2 \in \mathbb{C}^{n-\kappa}$. Then, we have

$$
\begin{aligned}
& \operatorname{rank} G'(x^*, 0) = 2n - \kappa \\
\Leftrightarrow\ & G'(x^*, 0) \begin{bmatrix} V_1 \cdot w_1 \\ w_2 \end{bmatrix} = 0 \text{ implies } w = 0, \\
\Leftrightarrow\ & F''(x^*) \begin{bmatrix} V_1 \cdot \boldsymbol{\lambda}_1 & V_1 \cdot w_1 \end{bmatrix} + F'(x^*) \cdot V_2 \cdot w_2 = 0 \text{ implies } w = 0, \\
\Leftrightarrow\ & A \cdot w = 0 \text{ implies } w = 0, \\
\Leftrightarrow\ & \operatorname{rank} A = n.
\end{aligned}
$$

Without loss of generality assume $\lambda_1 \neq 0$, then we prove that A is of full rank for almost all choices of bases $\mathrm{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\boldsymbol{\lambda}_1 = (\lambda_1, \ldots, \lambda_\kappa)^\top \in \mathbb{C}^\kappa$ is equivalent to the matrix

$$
C = \begin{bmatrix} F''(x^*)(v_1', v_1') & \cdots & F''(x^*)(v_\kappa, v_1') & F'(x^*) \cdot V_2 \end{bmatrix}
$$

is of full rank for almost all choices of bases $\mathrm{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$, where $v_1' = \sum_{i=1}^{\kappa} \lambda_i v_i \in \ker F'(x^*)$. This is true from the following equivalent statements:

$$
\begin{aligned}
& \mathrm{rank}A = n \\
\Leftrightarrow\ & \det\left[\sum_{i=1}^{\kappa}\lambda_i F''(x^*)(v_1,v_i) \quad \cdots \quad \sum_{i=1}^{\kappa}\lambda_i F''(x^*)(v_\kappa,v_i) \quad F'(x^*)\cdot V_2\right] \neq 0, \\
\Leftrightarrow\ & \det\left[F''(x^*)(v_1,\Lambda) \quad \cdots \quad F''(x^*)(v_\kappa,\Lambda) \quad F'(x^*)\cdot V_2\right] \neq 0, \\
\Leftrightarrow\ & \det\left[F''(x^*)(\Lambda,\Lambda) \quad F''(x^*)(v_2,\Lambda) \quad \cdots \quad F''(x^*)(v_\kappa,\Lambda) \quad F'(x^*)\cdot V_2\right] \neq 0, \\
\Leftrightarrow\ & \det\left[F''(x^*)(v_1',v_1') \quad F''(x^*)(v_2,v_1') \quad \cdots \quad F''(x^*)(v_\kappa,v_1') \quad F'(x^*)\cdot V_2\right] \neq 0, \\
\Leftrightarrow\ & \mathrm{rank}C = n
\end{aligned}
$$

where $\Lambda = \sum_{i=1}^{\kappa}\lambda_i v_i$. Denote the set of $n \times n$ unitary matrices by $\mathcal{U}(n)$. Note that $\mathcal{U}(n)$ is a manifold. Suppose $\mathrm{rank}B = n$ with probability one holds on $\mathcal{U}(n)$. Assume that for the same V_2 used in B,

$$
C = \left[F''(x^*)(v_1,v_1) \quad \cdots \quad F''(x^*)(v_\kappa,v_1) \quad F'(x^*)\cdot V_2\right]
$$

is not of full rank, then there exist $u \in \mathbb{C}^\kappa, w \in \mathbb{C}^{n-\kappa}$ ($u \neq 0$ or $w \neq 0$), such that

$$
u_1 F''(x^*)(v_1,v_1) + \cdots + u_\kappa F''(x^*)(v_\kappa,v_1) + F'(x^*)\cdot V_2 \cdot w = 0.
$$

If $u_2 = \cdots = u_\kappa = 0$, then B is not of full rank (normalizing v_1 by $\frac{v_1}{\|v_1\|}$ if needed), which is a contradiction. Without loss of generality, we assume $u_2 \neq 0$. Let $v_2' = u_1 v_1 + \cdots + u_\kappa v_\kappa$, then

$$
F''(x^*)(v_1,v_2') + F'(x^*)\cdot V_2 \cdot w = 0, \tag{3.2}
$$

where $\{v_1, v_2', v_3, \ldots, v_\kappa\}$ is a basis of $\ker F'(x^*)$. Clearly, $\{v_1 + v_2', v_1 - v_2', v_3, \ldots, v_\kappa\}$ is also a basis of $\ker F'(x^*)$. Applying the Gram-Schmidt process, we may assume that

$\{v_1 + v_2', v_1 - v_2', v_3, \ldots, v_\kappa, V_2\}$ is an orthonormal basis satisfying

$$
\begin{aligned}
& B(v_1 + v_2', v_1 - v_2', v_3, \ldots, v_\kappa) \\
= & \begin{bmatrix} F''(x^*)(v_1 + v_2', v_1 + v_2'), & F''(x^*)(v_1 - v_2', v_1 - v_2'), & \cdots \end{bmatrix} \\
= & \begin{bmatrix} F''(x^*)(v_1 + v_2', v_1 + v_2'), & F''(x^*)(v_1 - v_2', v_1 - v_2'), & \cdots \end{bmatrix} \\
= & \begin{bmatrix} F''(x^*)(v_1, v_1) + F''(x^*)(v_2', v_2') & F''(x^*)(v_1, v_1) + F''(x^*)(v_2', v_2') & \cdots \\ +2F''(x^*)(v_1, v_2') & -2F''(x^*)(v_1, v_2') & \end{bmatrix}.
\end{aligned}
$$

Note that even after the Gram-Schmidt process the directions of first two vectors $v_1 + v_2'$ and $v_1 - v_2'$ are not changed. Then, according to equation (3.2), $B(v_1 + v_2', v_1 - v_2', v_3, \ldots, v_\kappa)$ is not of full rank. This is a contradiction. These statements hold for any permutation of $\{1, \ldots, \kappa\}$.

For the "if" direction, we suppose that C is of full rank with probability one for almost all choices of bases $\mathrm{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$. Also, assume that the desired claim is not true. That is, there are a set of orthonormal vectors $\{\tilde{v}_1, \ldots, \tilde{v}_\kappa, \tilde{V}_2\} \in \mathcal{U}(n)$ satisfying that $\mathrm{span}_{\mathbb{C}}\{\tilde{v}_1, \ldots, \tilde{v}_\kappa\} = \ker F'(x^*)$ and a nontrivial open neighborhood $\tilde{U} \subset \mathcal{U}(n)$ of $\{\tilde{v}_1, \ldots, \tilde{v}_\kappa, \tilde{V}_2\}$ such that $\det B = 0$ for all points in $\tilde{U}$. Then, the identity theorem implies that $\det B = 0$ for any point in $\mathcal{U}(n)$. Let $\{v_1, \ldots, v_\kappa\}$ be a basis for $\ker F(x^*)$ with $\det C \neq 0$ and $W \in \mathbb{C}^{\kappa \times \kappa}$ be an invertible transformation such that

$$
\tilde{v}_i = \sum_{j=1}^{\kappa} W_{ij} v_j \quad \text{for } i = 1, \ldots, \kappa.
$$

If we express $\det B$ in terms of $v_1, \ldots, v_\kappa$, then $\det B$ is a (homogeneous) polynomial in variables W_{ij} for $i, j = 1, \ldots, \kappa$ of degree 2κ. We show that $\det B$ is not a zero polynomial,

leading to a contradiction. It is enough to find one nonzero term of $\det B$. Note that

$$
\begin{aligned}
& \det B \\
= \; & \det \left[F''(x^*)(\sum_{j=1}^{\kappa} W_{1j} v_j, \sum_{j=1}^{\kappa} W_{1j} v_j) \quad \cdots \quad F''(x^*)(\sum_{j=1}^{\kappa} W_{\kappa j} v_j, \sum_{j=1}^{\kappa} W_{\kappa j} v_j) \quad F'(x^*) \cdot V_2 \right] \\
= \; & W_{11}^2 W_{21} W_{22} \cdots W_{\kappa 1} W_{\kappa\kappa} \\
& \cdot \det \left[F''(x^*)(v_1, v_1) \quad 2F''(x^*)(v_1, v_2) \quad \cdots \quad 2F''(x^*)(v_1, v_\kappa) \quad F(x^*) \cdot V_2 \right] \\
& + \text{(other terms)}.
\end{aligned}
$$

Since we have $\det C \neq 0$ as a nonzero coefficient of $W_{11}^2 W_{21} W_{22} \cdots W_{\kappa 1} W_{\kappa\kappa}$ term, $\det B$ is not a zero polynomial. This is a contradiction. Consequently, we get $\mathrm{rank} B = n$ for almost all choices of orthonormal bases $\{v_1, \ldots, v_\kappa, V_2\}$ such that $\mathrm{im}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\mathrm{im}\, V_2 = \{\ker F'(x^*)\}^\perp$. □

The above characterization directly gives us the definition of simple multiple roots.

Definition 30. A point $x^* \in \mathbb{C}^n$ is a *simple multiple* root of a polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$, if

1. x^* is an isolated root of F,
2. $\dim \ker F'(x^*) = \kappa > 0$,
3. Let $\{v_1, \ldots, v_\kappa\}$ be a random orthonormal basis of $\ker F'(x^*)$, then with probability one, the linear operator

$$\mathcal{A} = F'(x^*) + \frac{F''(x^*)}{2}(v_1, \Pi_{v_1} \cdot) + \cdots + \frac{F''(x^*)}{2}(v_\kappa, \Pi_{v_\kappa} \cdot)$$

is invertible, where Π_{v_i} is the Hermitian projection to $\mathrm{span}_{\mathbb{C}}\{v_i\}$.

In fact, condition 3 is equivalent to the matrix

$$B = \begin{bmatrix} F''(x^*)(v_1, v_1) & \cdots & F''(x^*)(v_\kappa, v_\kappa) & F'(x^*) \cdot V_2 \end{bmatrix} \in \mathbb{C}^{n \times n}$$

being full rank, where $V_2 \in \mathbb{C}^{n \times (n-\kappa)}$ satisfying $\{v_1, \ldots, v_\kappa, V_2\}$ is an orthonormal basis with $\operatorname{im}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\operatorname{im} V_2 = \{\ker F'(x^*)\}^\perp$. The random choice of orthonormal basis makes it easy to select an invertible matrix B. Using Theorem 29 we can construct a suitable nonsingular linear operator $\mathcal{A}$ to prove the statements in the following sections. The next example shows how to get such $\mathcal{A}$.

Example 31. The system

$$F(x, y, z) = \begin{bmatrix} x^3 - yz \\ y^3 - xz \\ z^3 - xy \end{bmatrix}$$

is suggested in [Stu02], and the deflation process terminates after one step. If we let x^* be the origin, then the system has a simple multiple root x^* with $\kappa = 3, \rho = 4$ and $\mu = 11$ based on the data in [DZ05a]. If we consider $v_1 = (-\frac{1}{3}, \frac{2}{3}, \frac{2}{3}), v_2 = (\frac{2}{3}, -\frac{1}{3}, \frac{2}{3})$ and $v_3 = (\frac{2}{3}, \frac{2}{3}, -\frac{1}{3})$, then they form an orthonormal basis for $\ker F'(x^*)$. Also, we check that the matrix

$$B = F'(x^*) + \sum_{i=1}^{3} \frac{F''(x^*)}{2}(v_i, \Pi_{v_i} \cdot) = \begin{bmatrix} -\frac{2}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & -\frac{2}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & -\frac{2}{9} \end{bmatrix}$$

is invertible.

We now extend the setting to analytic systems. As pointed out in [DLZ11, Corollary 3], for any analytic system with an isolated zero, the system has the same multiplicity structure as the truncated polynomial system of its Taylor's series up to order depth at the common zero. Therefore, it is straightforward to generalize Theorem 27 and Theorem 29 to analytic systems with isolated zeros. The following example illustrates the characterization

for simple multiple roots of analytic systems.

Example 32. Let

$$F(x,y,z) = \begin{bmatrix} (y-z)^3 + (-x-y-z)\sin(x-z) \\ (x-z)^3 - (y-z)\sin(x+y+z) \\ (-x-y-z)^3 + (x-z)\sin(y-z) \end{bmatrix}$$

which is equivalent to the system $[u^3 + w\sin(v), v^3 + u\sin(w), w^3 + v\sin(u)]^\top$ with $u = y - z$, $v = x - z$ and $w = -u - v - w$ in [DLZ11, Example 3]. This system has a zero $x^* = (0,0,0)$ of $\kappa = 3, \rho = 4$ and $\mu = 11$. If we consider $v_1 = (1,0,0), v_2 = (0,1,0)$ and $v_3 = (0,0,1)$, then these vectors form an orthonormal basis for $\ker F'(x^*)$. Also, we check that the deflated system

$$G(x,y,z) = \begin{bmatrix} F \\ -2\sin(y-z) + (-x-y-z)\cos(x-z) + 3(y-z)^2 \\ 3(x-z)^2 - 2(y-z)\cos(x+y+z) - \sin(x+y+z) \\ -6(-x-y-z)^2 + \sin(y-z) + (x-z)\cos(y-z) \end{bmatrix}$$

with the vector $\boldsymbol{\lambda} = (1,1,0)^\top$ and its Jacobian matrix

$$G'(x^*) = \begin{bmatrix} & F'(x^*) & \\ -3 & -1 & 1 \\ -1 & -3 & 1 \\ 1 & 1 & -2 \end{bmatrix}$$

is of full rank three, which means that the deflation process terminates by one step. On the

other hand, we check the the matrix

$$B = F'(x^*) + \sum_{i=1}^{3} \frac{F''(x^*)}{2}(v_i, \Pi_{v_i}\cdot) = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

is invertible.

3.3 Lemmas

In this section, we suggest and prove inequalities that will be used to derive the separation bound. The desired inequalities provide a lower bound of a distance between the given multiple root and another zero of the system, and the bound only depends on the given system and its multiple root.

Since we extend Theorems 27 and 29 to the case of analytic systems, we deal with a square analytic system from now on. For an analytic system $F : \mathbb{C}^n \to \mathbb{C}^n$ and its simple multiple root x^*, we assume that $\dim \ker F'(x^*) = \kappa$. For a randomly chosen orthonormal basis $\{v_1, \ldots, v_\kappa, V_2\}$ with $\operatorname{im}\{v_1, \ldots, v_\kappa\} = \ker F'(x^*)$ and $\operatorname{im} V_2 = \{\ker F'(x^*)\}^\perp$, define an operator

$$\mathcal{A} = F(x^*) + \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(v_i, \Pi_{v_i}\cdot)$$

where Π_{v_i} is a Hermitian projection onto $\operatorname{span}_{\mathbb{C}}\{v_i\}$. In the case of simple multiple zeros, we know that the operator $\mathcal{A}$ is invertible. Moreover, we introduce the parameter γ_κ which depends on F and x^* such that

$$\gamma_\kappa(F, x^*) = \max\left\{1, \sup_{k\geq 2}\left\|\mathcal{A}^{-1}\frac{F^{(k)}(x^*)}{k!}\right\|^{\frac{1}{k-1}}\right\}. \tag{3.3}$$

We use γ_κ if F and x^* are obvious in the context. Finally, we employ a constant d obtained

from the smallest positive real root of

$$\sqrt{1-d^2} - (\kappa+1)\kappa d\sqrt{1-d^2} - \kappa d^2 - d = 0.$$

The same setting will be used in §3.4 also.

Let $a, b \in \mathbb{C}^n$ be any two vectors. Then, we know that the angle between a and b can be defined by

$$d_P(a,b) = \arccos \frac{|\langle a, b\rangle|}{\|a\|\|b\|}.$$

For a simple multiple root x^* and an arbitrary point $y \in \mathbb{C}^n$, we define the direction vector $w = y - x$ between x and y. Using orthonormal vectors $\{v_1, \ldots, v_\kappa\}$ obtained from Definition 30, we can represent w as

$$w = y - x = \hat{w} + \alpha_1 v_1 + \cdots + \alpha_\kappa v_\kappa.$$

From the trigonometric definition of the angle, we define $\varphi = d_P(w - \hat{w}, w)$ and $\varphi_i = d_P(v_i, w)$ for $i = 1, \ldots, \kappa$. Then, we have

$$\|\hat{w}\| = \|w\| \sin\varphi, \quad |\alpha_i| = \|w\| \cos\varphi_i \quad \text{and} \quad \|w - \alpha_i v_i\| = \|w\| \sin\varphi_i \quad \text{for } i = 1, \ldots, \kappa.$$

We now provide a series of lemmas. The main idea is constructing a lower bound of $\|w\|$ using the angle φ between x^* and y. In the next section, we consider y as another root of F and this lower bound of $\|w\|$ gives us the separation bound. The Taylor expansion will be the trick to derive the lower bound of $\|w\|$. We define a constant angle θ such that $\sin\theta = \frac{d}{\gamma_\kappa}$. As the first step, we deal with the case when φ is big, i.e. $\varphi \geq \theta$.

Lemma 33. Assume that $\gamma_\kappa \|w\| \leq \frac{1}{2}$ and $\varphi \geq \theta$ for a fixed y. Then, we have

$$\left\|\mathcal{A}^{-1}F(y)\right\| \geq \|w\| \sin\theta - 2\gamma_\kappa \|w\|^2.$$

Proof. Applying the Taylor expansion on $F(y)$ at x^* gives us that

$$\begin{aligned} F(y) &= F(x^*) + F'(x^*)w + \sum_{k\geq 2} \frac{F^{(k)}(x^*)w^k}{k!} \\ &= \mathcal{A}w - \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(v_i, \Pi_{v_i} w) + \sum_{k\geq 2} \frac{F^{(k)}(x^*)w^k}{k!} \end{aligned} \tag{3.4}$$

We observe that

$$\mathcal{A}^{-1} \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(v_i, \Pi_{v_i} w) = \mathcal{A}^{-1} \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(v_i, \alpha_i v_i) = \sum_{i=1}^{\kappa} \alpha_i \mathcal{A}^{-1}\mathcal{A} v_i = \sum_{i=1}^{\kappa} \alpha_i v_i$$

because we have vectors v_i which are orthonormal. Therefore, multiplying $\mathcal{A}^{-1}$ on both sides of (3.4) gives us $\hat{w}$ on the right hand side. When we solve for $\hat{w}$, we obtain

$$\hat{w} = \mathcal{A}^{-1}F(y) - \sum_{k\geq 2} \mathcal{A}^{-1} \frac{F^{(k)}(x^*)w^k}{k!}.$$

We combine the facts that $\|\hat{w}\| = \|w\| \sin\varphi$ and $\varphi \geq \theta$, leading to the conclusion that

$$\begin{aligned} \|w\| \sin\theta \leq \|w\| \sin\varphi &= \|\hat{w}\| \\ &\leq \left\|\mathcal{A}^{-1}F(y)\right\| + \sum_{k\geq 2} \left\| \mathcal{A}^{-1} \frac{F^{(k)}(x^*)w^k}{k!} \right\| \\ &\leq \left\|\mathcal{A}^{-1}F(y)\right\| + \sum_{k\geq 2} \gamma_\kappa^{k-1} \|w\|^k \\ &\leq \left\|\mathcal{A}^{-1}F(y)\right\| + \gamma_\kappa \|w\|^2 \sum_{k\geq 0} \left(\frac{1}{2}\right)^k \\ &= \left\|\mathcal{A}^{-1}F(y)\right\| + 2\gamma_\kappa \|w\|^2. \end{aligned}$$

We use the assumption that $\gamma_\kappa \|w\| \leq \frac{1}{2}$ in order to get the last inequality. □

In the case φ is small (that is, $\varphi \leq \theta$), we need to define a supplementary operator as in the following lemma:

Lemma 34. Define the operator

$$\mathcal{A}_{(\alpha_1,\dots,\alpha_\kappa)} = F'(x) + \sum_{i=1}^{\kappa} F''(x)(\alpha_i v_i, \Pi_{v_i}\cdot)$$

where $\alpha_i \neq 0$ for all i. Then, $\mathcal{A}_{(\alpha_1,\dots,\alpha_\kappa)}$ is nonsingular and

$$\|\mathcal{A}^{-1}_{(\alpha_1,\dots,\alpha_\kappa)}\mathcal{A}_{(\beta_1,\dots,\beta_\kappa)}\| = \max\left\{1, \left|\frac{\beta_1}{\alpha_1}\right|, \dots, \left|\frac{\beta_\kappa}{\alpha_\kappa}\right|\right\}.$$

Proof. Without loss of generality, take any vector $z \in \mathrm{span}_{\mathbb{C}}\{v_1\}$. Then, we get

$$\mathcal{A}^{-1}_{(\alpha_1,\dots,\alpha_\kappa)}\mathcal{A}_{(\beta_1,\dots,\beta_\kappa)}z = \frac{\beta_1}{\alpha_1}z.$$

On the other hand, if we take a vector $z \in \mathrm{span}_{\mathbb{C}}\{v_1,\dots,v_\kappa\}^{\perp}$, then we have that $\mathcal{A}^{-1}_{(\alpha_1,\dots,\alpha_\kappa)}\mathcal{A}_{(\beta_1,\dots,\beta_\kappa)}z = z$. Hence, we get the desired result. □

Assume that we have $\mathcal{A}_{(\alpha_1,\dots,\alpha_\kappa)}$ such that all α_i are nonzero. It is possible because $\{v_1,\dots,v_\kappa\}$ is chosen generically. We apply this operator to prove the next inequality which is used to deal with the small angle case.

Lemma 35. Assume that $\gamma_\kappa\|w\| \leq \frac{1}{2}$. Then,

$$\begin{aligned}\left\|\mathcal{A}^{-1}F(y)\right\| \geq \|w\| &\min_{i=1,\dots,\kappa}|\alpha_i| \\ &- (\kappa+1)\gamma_\kappa\|w\|^2\sum_{i=1}^{\kappa}\sin\varphi_i\cos\varphi_i - \gamma_\kappa\|w\|^2\sum_{i=1}^{\kappa}\sin^2\varphi_i - 2\gamma_\kappa^2\|w\|^3.\end{aligned}$$

Proof. For brevity, we denote $\mathcal{A}_{\frac{\alpha}{2}} = \mathcal{A}_{(\frac{\alpha_1}{2},\dots,\frac{\alpha_\kappa}{2})}$, $\mathcal{A}_{\frac{1}{2}} = \mathcal{A}_{(\frac{1}{2},\dots,\frac{1}{2})}$ and $\hat{\mathcal{A}}^{-1} = \mathcal{A}^{-1}_{\frac{\alpha}{2}}\mathcal{A}_{\frac{1}{2}}\mathcal{A}^{-1}_{\frac{1}{2}}$.

We apply the Taylor expansion of f centered at x^*. Then,

$$\begin{aligned}
F(y) &= F(x^*) + F'(x^*)w + \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(\alpha_i v_i, \alpha_i v_i) \\
&\qquad - \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(\alpha_i v_i, \alpha_i v_i) + \sum_{k\geq 2} \frac{F^{(k)}(x^*)w^k}{k!} \\
&= \mathcal{A}_{\frac{\alpha}{2}} w - \sum_{i=1}^{\kappa} \frac{F''(x^*)}{2}(\alpha_i v_i, \alpha_i v_i) + \frac{F''(x^*)}{2}(w, w) + \sum_{k\geq 3} \frac{F^{(k)}(x^*)w^k}{k!} \\
&= \mathcal{A}_{\frac{\alpha}{2}} w + \sum_{i=1}^{\kappa} F''(x^*)(\hat{w}, \alpha_i v_i) + \sum_{i\neq j}^{\kappa} \frac{F''(x^*)}{2}(\alpha_i v_i, \alpha_j v_j) \\
&\qquad + \frac{F''(x^*)}{2}(\hat{w}, \hat{w}) + \sum_{k\geq 3} \frac{F^{(k)}(x^*)w^k}{k!}
\end{aligned}$$

By genericity, we may assume that $\mathcal{A}_{\frac{\alpha}{2}}$ is invertible : $\mathcal{A}_{\frac{1}{2}}$ should also be invertible. Multiplying $\hat{\mathcal{A}}^{-1}$ on both sides, we get

$$\begin{aligned}
\hat{\mathcal{A}}^{-1}F(y) = w + \hat{\mathcal{A}}^{-1} \sum_{i=1}^{\kappa} F''(x^*)(\hat{w}, \alpha_i v_i) + \hat{\mathcal{A}}^{-1} \sum_{i\neq j}^{\kappa} \frac{F''(x^*)}{2}(\alpha_i v_i, \alpha_j v_j) \\
+ \hat{\mathcal{A}}^{-1} \frac{F''(x^*)}{2}(\hat{w}, \hat{w}) + \hat{\mathcal{A}}^{-1} \sum_{k\geq 3} \frac{F^{(k)}(x^*)w^k}{k!}. \quad (3.5)
\end{aligned}$$

Now, we derive the desired inequality. We first recall that $\mathcal{A}_{\frac{1}{2}} = \mathcal{A}$ and

$$|\alpha_i| = \langle w, v_i \rangle \leq \|w\| \leq \gamma_\kappa \|w\| \leq \frac{1}{2}.$$

Also, we know that $\|\mathcal{A}_{\frac{\alpha}{2}}^{-1}\mathcal{A}_{\frac{1}{2}}\| = \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|}$ by Lemma 34. We combine these and subtract all terms in the right hand side of the equation (3.5) except for w. Then, applying the triangle inequality, we have

$$
\begin{aligned}
\|w\| \leq & \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \left\|\mathcal{A}^{-1}F(y)\right\| + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \sum_{i=1}^{\kappa} \left\|\mathcal{A}^{-1}F''(x^*)\right\| \|\hat{w}\| \|\alpha_i v_i\| \\
& + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \sum_{i\neq j}^{\kappa} \left\| \frac{\mathcal{A}^{-1}F''(x^*)}{2} \right\| \|\alpha_i v_i\| \|\alpha_j v_j\| \\
& + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \left\| \frac{\mathcal{A}^{-1}F''(x^*)}{2} \right\| \|\hat{w}\|^2 + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \sum_{k\geq 3} \left\| \frac{\mathcal{A}^{-1}F^{(k)}(x^*)}{k!} \right\| \|w\|^k \\
\leq & \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \left\|\mathcal{A}^{-1}F(y)\right\| + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \sum_{i=1}^{\kappa} \left\|\mathcal{A}^{-1}F''(x^*)\right\| \|w - \alpha_i v_i\| \|\alpha_i v_i\| \\
& + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} (\kappa - 1) \sum_{i=1}^{\kappa} \left\| \frac{\mathcal{A}^{-1}F''(x^*)}{2} \right\| \|w - \alpha_i v_i\| \|\alpha_i v_i\| \\
& + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \left\| \frac{\mathcal{A}^{-1}F''(x^*)}{2} \right\| \|\hat{w}\|^2 + \max_{i=1,\ldots,\kappa} \frac{1}{|\alpha_i|} \sum_{k\geq 3} \left\| \frac{\mathcal{A}^{-1}F^{(k)}(x^*)}{k!} \right\| \|w\|^k.
\end{aligned}
$$

The last inequality is obtained from $\|\hat{w}\| \leq \|w - \alpha_i v_i\|$ for any i and $\|\alpha_i v_i\| \leq \|w - \alpha_j v_j\|$ whenever $i \neq j$. We observe that $|\alpha_i| = \|w\| \cos\varphi_i$, $\|w - \alpha_i v_i\| = \|w\| \sin\varphi_i$ and $\|\hat{w}\| = \|w\| \sin\varphi$. Then, we have

$$
\cos^2\varphi = \sum_{i=1}^{\kappa} \cos^2\varphi_i, \quad \text{and} \quad \sin^2\varphi \leq \sum_{i=1}^{\kappa} \sin^2\varphi_i
$$

from basic trigonometric equalities. Simplifying the inequality based on the above trigonometric expressions, we get

$$
\begin{aligned}
\|w\| \min_{i=1,\ldots,\kappa} |\alpha_i| \leq \left\|\mathcal{A}^{-1}F(y)\right\| + (\kappa+1)\gamma_\kappa \|w\|^2 \sum_{i=1}^{\kappa} \sin\varphi_i \cos\varphi_i \\
+ \gamma_\kappa \|w\|^2 \sum_{i=1}^{\kappa} \sin^2\varphi_i + 2\gamma_\kappa^2 \|w\|^3
\end{aligned}
$$

The last term of the inequality attained from the assumption that $\gamma_\kappa \|w\| \leq \frac{1}{2}$. Solving the inequality for $\|\mathcal{A}^{-1}F(y)\|$ derives the desired result. □

The next lemma combines the results from Lemmas 33 and 35.

Lemma 36. For w and γ_κ satisfying $\gamma_\kappa\|w\| \le \frac{1}{2}$, we have

1. $\|\mathcal{A}^{-1}F(y)\| \ge 2\gamma_\kappa\|w\|\left(\frac{\sin\theta}{2\gamma_\kappa} - \|w\|\right)$ if $\varphi \ge \theta$.
2. $\|\mathcal{A}^{-1}F(y)\| \ge 2\gamma_\kappa^2\|w\|^2\left(\frac{\sin\theta}{2\gamma_\kappa} - \|w\|\right)$ if $\varphi \le \theta$.

Proof. For the case of $\varphi \ge \theta$, we get the result from Lemma 33.

Now, assume that we have $\varphi \le \theta$. Without loss of generality, we fix $\varphi_2, \ldots, \varphi_\kappa$. Then,

$$\min_{i=1,\ldots,\kappa}(\cos\varphi_i) - \gamma_\kappa(\kappa+1)\sum_{i=1}^{\kappa}\sin\varphi_i\cos\varphi_i - \gamma_\kappa\sum_{i=1}^{\kappa}\sin^2\varphi_i$$

is a (univariate) decreasing function for $\varphi_1 \in [0, \frac{\pi}{4}]$. Since it is decreasing for each φ_i, the univariate function

$$g(\varphi) := \cos\varphi - \gamma_\kappa(\kappa+1)\kappa\sin\varphi\cos\varphi - \gamma_\kappa\kappa\sin^2\varphi$$

is also decreasing for $\varphi \in [0, \frac{\pi}{4}]$. This means that if $\theta \in [0, \frac{\pi}{4}]$, then by Lemma 35 we have

$$\left\|\mathcal{A}^{-1}F(y)\right\| \ge 2\gamma_\kappa^2\|w\|^2\left(\frac{g(\theta)}{2\gamma_\kappa^2} - \|w\|\right)$$

because we assume $\varphi \le \theta$. Therefore, it is enough to show that

$$\frac{g(\theta)}{2\gamma_\kappa^2} \ge \frac{\sin\theta}{2\gamma_\kappa}. \tag{3.6}$$

Using the definition of θ, we define

$$h(d,\gamma_\kappa) := \sqrt{1-\frac{d^2}{\gamma_\kappa^2}} - (\kappa+1)\kappa d\sqrt{1-\frac{d^2}{\gamma_\kappa^2}} - \kappa^2\frac{d^2}{\gamma_\kappa} - d$$

which is obtained from inequality (3.6). We want to show that $h(d,\gamma_\kappa) \ge 0$ if $\gamma_\kappa \ge 1$. Note that if we consider h as a univariate function of γ_κ, then $h(\gamma_\kappa)$ is increasing when $\gamma_\kappa \ge 1$ for any $d \le \frac{1}{2}$. Thus, it is enough to check the case of $\gamma_\kappa = 1$ that $h(d,1) \ge 0$. It is clear

since we let d be the root of $h(d, 1) = 0$. □

3.4 Main results

In this section, we show our main results. We use the setting of §3.3. The first theorem provides the separation bound for an exact analytic system and its simple multiple root.

Theorem 37. Let x^* be a simple multiple root of the given square analytic system $F : \mathbb{C}^n \to \mathbb{C}^n$. Then, if we let y be another root of F, then

$$\|y - x^*\| \geq \frac{d}{2\gamma_\kappa(F, x^*)^2}.$$

Proof. Define $w = y - x^*$. Noting that $F(y) = 0$, if we have $\gamma_\kappa(F, x^*)\|w\| \leq \frac{1}{2}$, then we get

$$\|w\| \geq \frac{\sin\theta}{2\gamma_\kappa(F, x^*)} = \frac{d}{2\gamma_\kappa(F, x^*)^2}$$

from Lemma 36. On the other hand, if $\gamma_\kappa(F, x^*)\|w\| \geq \frac{1}{2}$, the claim follows from the fact that $d < 1$ and $\gamma_\kappa(F, x^*) \geq 1$. □

The next theorem describes the behavior of points close to the multiple root. That is, when a point $y \in \mathbb{C}^n$ different from x^* is contained in some neighborhood of the multiple root, the value of $F(y)$ is strictly greater than 0.

Theorem 38. Let x^* be a simple multiple root of the given square analytic system $F : \mathbb{C}^n \to \mathbb{C}^n$. Then, for any $y \in \mathbb{C}^n$ satisfying $\|y - x^*\| \leq \frac{d}{4\gamma_\kappa(F,x^*)^2}$, we have

$$\|F(y)\| \geq \frac{d\|y - x^*\|^2}{2\,\|\mathcal{A}^{-1}\|}.$$

Proof. Define $w = y - x^*$ as above. Since we assume that $\|w\| \leq \frac{d}{4\gamma_\kappa(F,x^*)^2}$, we have

$\|w\| \leq \frac{\sin\theta}{4\gamma_\kappa(F,x^*)}$. Then, applying Lemma 36, we obtain

$$\begin{aligned}\left\|\mathcal{A}^{-1}\right\| \|F(y)\| \geq \left\|\mathcal{A}^{-1}F(y)\right\| &\geq 2\gamma_\kappa(F,x^*)^2\|w\|^2\left(\frac{\sin\theta}{2\gamma_\kappa(F,x^*)} - \|w\|\right) \\ &\geq 2\gamma_\kappa(F,x^*)^2\|w\|^2\frac{\sin\theta}{4\gamma_\kappa(F,x^*)} = \frac{d\|w\|^2}{2}\end{aligned}$$

which proves the claim. □

From now on, we deal with an analytic system G close to F. In order to depict such a system, we need the 'local distance around x' between two systems F and G, i.e. for $R > 0$, we define

$$d_R(F,G) = \max_{\|y-x\|\leq R} \|F(y) - G(y)\|.$$

Using Theorem 38, we have a cluster of roots of G which corresponds to x^* of F.

Theorem 39. Let x^* be a simple multiple root with the multiplicity μ of the given square analytic system $F : \mathbb{C}^n \to \mathbb{C}^n$. Let $G : \mathbb{C}^n \to \mathbb{C}^n$ be another analytic system. Let R be a positive number satisfies that $0 < R \leq \frac{d}{4\gamma_\kappa(F,x^*)^2}$ and $d_R(F,G) < \frac{dR^2}{2\|\mathcal{A}^{-1}\|}$. Then, there are μ zeros (up to multiplicity) of G in $B(x^*,R)$.

Proof. If we have y with $\|y - x^*\| = R$, then

$$\|F(y) - G(y)\| \leq d_R(F,G) < \frac{dR^2}{2\,\|\mathcal{A}^{-1}\|} \leq \|F(y)\|$$

because of Theorem 38. Therefore, F and G have the same number of zeros (up to multiplicity) in $B(x^*,R)$ by Rouché's theorem. By Theorem 37, we know that F has only one solution x^* with the multiplicity μ in $B(x^*,R)$. Therefore, G has μ zeros in $B(x^*,R)$. □

We introduce an application of Theorem 39. Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be an analytic system, x be a point in $\mathbb{C}^n$. We may consider the system F as a system having a simple multiple zero and a point x as a point approxiating the simple multiple zero of F. Let $\{v_1,\ldots,v_\kappa\}$ be vectors in $\mathbb{C}^n$ such that for any $i,j = 1,\ldots,\kappa$, $\|v_i\| = 1$ and $\langle v_i, v_j\rangle = 0$ if $i \neq j$. We further

assume that $\mathrm{rank} F'(x)|_{\mathrm{span}_{\mathbb{C}}\{v_1,\ldots,v_\kappa\}^\perp} = n-\kappa$ and $F''(x)(v_i, v_i) \notin \mathrm{im}\, F'(x)|_{\mathrm{span}_{\mathbb{C}}\{v_1,\ldots,v_\kappa\}^\perp}$ for all $i = 1, \ldots, \kappa$.

Based on this setting, we define the linear operator $\mathcal{H} : \mathbb{C}^n \to \mathbb{C}^n$ by $\mathcal{H}(v_i) = F'(x)v_i$ and $\mathcal{H}(z) = 0$ if $z \in \mathrm{span}_{\mathbb{C}}\{v_1, \ldots, v_\kappa\}^\perp$. Then, we have a linear operator $\mathcal{A} - \mathcal{H}$ which is nonsingular. We define a new parameter $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$ using the operator $\mathcal{A} - \mathcal{H}$ in a way that

$$\gamma_\kappa(F, x, v_1, \ldots, v_\kappa) = \max\left\{1, \sup_{k\geq 2}\left\|(\mathcal{A}-\mathcal{H})^{-1}\frac{F^{(k)}(x)}{k!}\right\|^{\frac{1}{k-1}}\right\}. \tag{3.7}$$

Then, the following theorem is attained from an application of Theorem 39:

Theorem 40. Let x be a point and $F : \mathbb{C}^n \to \mathbb{C}^n$ be a square analytic system. Suppose that

$$\|F(x)\| + \|\mathcal{H}\|\frac{d}{4\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)^2} < \frac{d^3}{32\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)^4\,\|(\mathcal{A}-\mathcal{H})^{-1}\|},$$

where $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$ is defined by (3.7). Then, F has at least 2^κ zeros (up to multiplicity) in $B(x, \frac{d}{4\gamma_\kappa(F,x,v_1,\ldots,v_\kappa)^2})$.

Proof. We define the system

$$G(y) = F(y) - F(x) - \mathcal{H}(y - x). \tag{3.8}$$

This system is constructed to obtain the two properties that $G'(x) = F'(x) - \mathcal{H}$ and $G^{(k)}(x) = F^{(k)}(x)$ for all $k \geq 2$. Also, $\dim\ker G'(x) = \kappa$ and $\{v_1, \ldots, v_\kappa\}$ is an orthonormal basis of $\ker G'(x)$. From an observation that $G'(x) = F'(x)|_{\mathrm{span}_{\mathbb{C}}\{v_1,\ldots,v_\kappa\}^\perp}$, we obtain an invertible operator

$$G'(x) + \sum_{i=1}^{\kappa}\frac{G''(x)}{2}(v_i, \Pi_{v_i}\cdot).$$

This shows that G is a system to which we can apply Theorem 39, and applying Theorem

39 proves the claim. □

From the proof of the theorem, the $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$ defined (3.7) is equal to $\gamma_\kappa(G, x)$ for G defined in equation (3.8). Moreover, there are the multiplicity of x for G many zeros inside $B(x, \frac{d}{4\gamma_\kappa(F,x,v_1,\ldots,v_\kappa)^2})$. However, we point out that one only knows the lower bound (2^κ) of the number of zeros inside the ball because in an actual application, we don't know the exact multiplicity of x for G. One might obtain how many zeros are there by observing a combinatorial property of G or some numerical approaches.

Remark 41. In order to implement the results suggested in this section, we need to compute the value of $\gamma_\kappa(F, x)$ and $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$ which are defined in equations (3.3) and (3.7). One can observe that the definition of them are similar to $\gamma(F, x)$ in equation (2.5). Therefore, the way to calculate the value of $\gamma(F, x)$ can be used for $\gamma_\kappa(F, x)$ and $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$ also. It is well-known that computing the operator norm of tensors of order larger than two is NP-hard [HL13]. Therefore, we can use the supremum norm to compute an upper bound of $\gamma_\kappa(F, x)$ according to [Giu+07, Lemma B.2]. It is also possible to get an easily computable bound on the operator norm of tensors according to [FL18, Lemma 9.1]. For polynomial systems, the upper bound of $\gamma(F, x)$ is suggested in [HS12]. In the case of some special analytic systems, there are also ways to get such upper bounds. In [HL17], the way to bound $\gamma(F, x)$ is suggested for the system with solutions of linear ODEs with constant coefficients. Even more, Theorem 19 provides the method for the system with univariate analytic functions.

In the case of computing $\gamma_\kappa(F, x, v_1, \ldots, v_\kappa)$, we need to know κ and orthonormal vectors $v_1, \ldots, v_\kappa$. One way to obtain those is the singular value decomposition of $F'(x)$. After decompose $F'(x)$, we only take sufficiently large singular values. The number of such singular values is κ and their corresponding orthonormal vectors can be used as $v_1, \ldots, v_\kappa$.

We give an example describing the effectiveness of our results.

Example 42 (Example 3 continued)**.** The system

$$F(x,y,z) = \begin{bmatrix} x^3 - 3x^2y + 3xy^2 - y^3 - z^2 \\ z^3 - 3z^2x + 3zx^2 - x^3 - y^2 \\ y^3 - 3y^2z + 3yz^2 - z^3 - x^2 \end{bmatrix}$$

has a simple multiple zero at $x^* = (0,0,0)$ with $\kappa = 3$ and $\mu = 8$. We compute the upper bound of $\gamma_\kappa(F, x^*) \leq 11.25$ as suggested in [HS12]. Then, we get the separation bound $\frac{d}{2\gamma_\kappa(F,x^*)^2} \approx 0.0003$ which is better than the global separation bound $\ll 10^{-10}$ suggested in [EMT10].

We recall that in §2.4.1.2, the function `certifySolution` can not certify an approximation of x^* since x^* is singular. Theorem 40 can be applied to certify x^*, i.e. checking the existence of at least 2^κ approximating a singular root x^* of F in a given compact region. We solve the system F numerically using `Macaulay2` [GS] package `NumericalAlgebraicGeometry` [Ley11], and obtain 8 numerical roots around x^*. Let one of them be $x_1^* = (-3.4 \cdot 10^{-20} - 4.2 \cdot 10^{-20}i, -3.4 \cdot 10^{-20} - 4.2 \cdot 10^{-20}i, -3.4 \cdot 10^{-20} - 4.2 \cdot 10^{-20}i)$ and use bounds $1 \leq \gamma_\kappa(F, x_1^*, v_1, \ldots, v_\kappa) \leq 11.25$, then we have

$$\begin{aligned} \|F(x_1^*)\| + \|\mathcal{H}\|\frac{d}{4\gamma_\kappa(F, x_1^*, v_1, \ldots, v_\kappa)^2} &\leq 4.3 \times 10^{-21} \\ &< 3.4 \times 10^{-9} \leq \frac{d^3}{32\gamma_\kappa(F, x_1^*, v_1, \ldots, v_\kappa)^4 \|(\mathcal{A} - \mathcal{H})^{-1}\|}. \end{aligned}$$

Therefore, according to Theorem 40, we know that F has at least $2^3 = 8$ zeros (up to multiplicity) in the ball $B(x_1^*, \frac{d}{4\gamma_\kappa(F,x_1^*,v_1,\ldots,v_\kappa)^2})$. One can repeat the computation for other 7 numerical roots, and will get the same results. Thus, we certify all approximations around at the origin. In fact, the exact singular zero x^* of F belongs to the ball $B(x_1^*, \frac{d}{4\gamma_\kappa(F,x_1^*,v_1,\ldots,v_\kappa)^2})$ since $\|x_1^* - x^*\| \approx 1.4 \cdot 10^{-19} < \frac{d}{4\gamma_\kappa(F,x_1^*,v_1,\ldots,v_\kappa)^2} \approx 0.00015$.

REFERENCES

[BJM17] A. Benoit, Mioara J., and M. Mezzarobba. "Rigorous uniform approximation of D-finite functions using Chebyshev expansions". In: *Mathematics of Computation* 86.305 (2017), pp. 1303–1341.

[Bli+18] N. Bliss et al. "Monodromy Solver: Sequential and Parallel". In: *Proceedings of the 2018 ACM International Symposium on Symbolic and Algebraic Computation*. ISSAC '18. New York, NY, USA: ACM, 2018, pp. 87–94. ISBN: 978-1-4503-5550-6.

[Blu+12] L. Blum et al. *Complexity and real computation*. Springer Science & Business Media, 2012.

[Bou06] N. Bourbaki. *Algèbre commutative: Chapitres 8 et 9*. Reprint of the 1983 original. Springer Berlin Heidelberg, 2006. ISBN: 9783540339809.

[BLR15] S. Bozóki, T. L. Lee, and L. Rónyai. "Seven mutually touching infinite cylinders". In: *Comput. Geom.* 48.2 (2015), pp. 87–93.

[BT18] Paul Breiding and Sascha Timme. "HomotopyContinuation. jl: A package for homotopy continuation in Julia". In: *International Congress on Mathematical Software*. Springer. 2018, pp. 458–465.

[BLL19] M. Burr, K. Lee, and A. Leykin. "Effective certification of approximate solutions to systems of equations involving analytic functions". In: *Proceedings of the 2019 on International Symposium on Symbolic and Algebraic Computation*. 2019, pp. 267–274.

[CC90] D. V. Chudnovsky and G. V. Chudnovsky. "Computer algebra in the service of mathematical physics and number theory". In: *Computers in mathematics* 125 (1990), p. 109.

[CLO13] D. Cox, J. Little, and D. OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

[DLZ11] B. Dayton, T. Li, and Z. Zeng. "Multiple Zeros of Nonlinear Systems". In: *Mathematics of Computation* 80 (2011), pp. 2143–2168.

[DZ05a] B. Dayton and Z. Zeng. "Computing the multiplicity structure in solving polynomial systems". In: *Proceedings of the 2005 International Symposium*

on Symbolic and Algebraic Computation. Ed. by M. Kauers. Beijing, China: ACM, 2005, pp. 116–123. ISBN: 1-59593-095-7.

[DZ05b] Barry H Dayton and Zhonggang Zeng. "Computing the multiplicity structure in solving polynomial systems". In: *Proceedings of the 2005 international symposium on Symbolic and algebraic computation*. ACM. 2005, pp. 116–123.

[DS01] J. P. Dedieu and M. Shub. "On simple double zeros and badly conditioned zeros of analytic functions of n variables". In: *Mathematics of computation* (2001), pp. 319–327.

[Duf+] T. Duff et al. *MonodromySolver: a Macaulay2 package for solving polynomial systems via homotopy continuation and monodromy*. Available at `http://people.math.gatech.edu/~aleykin3/MonodromySolver`.

[Duf+19] T. Duff et al. "Solving polynomial systems via homotopy continuation and monodromy". In: *IMA Journal of Numerical Analysis* 39.3 (2019), pp. 1421–1446.

[EMT10] I. Z. Emiris, B. Mourrain, and E. P. Tsigaridas. "The DMM bound: Multivariate (aggregate) separation bounds". In: *Proceedings of the 2010 International Symposium on Symbolic and Algebraic Computation*. ACM. 2010, pp. 243–250.

[EV14] I. Z. Emiris and R. Vidunas. "Root Counts of Semi-mixed Systems, and an Application to Counting Nash Equilibria". In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*. ISSAC '14. Kobe, Japan: ACM, 2014, pp. 154–161. ISBN: 978-1-4503-2501-1.

[FL18] S. Friedland and L. H. Lim. "Nuclear norm of higher-order tensors". In: *Mathematics of Computation* 87 (2018), pp. 1255–1281.

[Giu+07] M. Giusti et al. "On Location and Approximation of Clusters of Zeros: case of Embedding Dimension One". In: *Foundations of Computational Mathematics* 7.1 (2007), pp. 1–58.

[GS] D. R. Grayson and M. E. Stillman. *Macaulay2, a software system for research in algebraic geometry*. Available at `http://www.math.uiuc.edu/Macaulay2/`.

[Hao+20] Z. Hao et al. "On isolation of simple multiple zeros and clusters of zeros of polynomial systems". In: *Mathematics of Computation* 89.322 (2020), pp. 879–909.

[HL17] J. D. Hauenstein and V. Levandovskyy. “Certifying solutions to square systems of polynomial-exponential equations”. In: *Journal of Symbolic Computation* 79 (2017), pp. 575–593.

[HS11] J. D. Hauenstein and F. Sottile. *alphaCertified: Software for certifying numerical solutions to polynomial equations*. Available at `math.tamu.edu/~sottile/research/stories/alphaCertified`. 2011.

[HS12] J. D. Hauenstein and F. Sottile. “Algorithm 921: alphaCertified: certifying solutions to polynomial systems”. In: *ACM Transactions on Mathematical Software (TOMS)* 38.4 (2012), p. 28.

[HL13] C. J. Hillar and L. H. Lim. “Most Tensor Problems Are NP-Hard”. In: *J. ACM* 60.6 (Nov. 2013), 45:1–45:39.

[Hoe99] J. van der Hoeven. “Fast evaluation of holonomic functions”. In: *Theoretical Computer Science* 210.1 (1999), pp. 199–215.

[Hoe03] J. van der Hoeven. *Majorants for formal power series*. Tech. rep. 2003-15. Université Paris-Sud, Orsay, France, 2003.

[Ked01] Kiran Kedlaya. “The algebraic closure of the power series field in positive characteristic”. In: *Proceedings of the American Mathematical Society* 129.12 (2001), pp. 3461–3470.

[Kra69] R. Krawczyk. “Newton-algorithmen zur bestimmung von nullstellen mit fehlerschranken”. In: *Computing* 4.3 (1969), pp. 187–201.

[Lan83] S. Lang. *Real analysis*. Second. Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1983, pp. xv+533. ISBN: 0-201-14179-5.

[Lee19] K. Lee. “Certifying approximate solutions to polynomial systems on `Macaulay2`”. In: *ACM Communications in Computer Algebra* 53.2 (2019), pp. 45–48.

[LLZ20] Kisun Lee, Nan Li, and Lihong Zhi. “On isolation of singular zeros of multivariate analytic systems”. In: *Journal of Symbolic Computation* (2020).

[LLT08] T.-L. Lee, T.-Y. Li, and C.-H. Tsai. “HOM4PS-2.0: A software package for solving polynomial systems by the polyhedral homotopy continuation method”. In: *Computing* 83.2-3 (2008), pp. 109–133.

[Ley11] A. Leykin. “Numerical algebraic geometry”. In: *Journal of Software for Algebra and Geometry* 3.1 (2011), pp. 5–10.

[LVZ06] A. Leykin, J. Verschelde, and A. Zhao. “Newton’s method with deflation for isolated singularities of polynomial systems”. In: *Theoretical Computer Science* 359.1-3 (2006), pp. 111–122.

[LZ12] N. Li and L. Zhi. “Computing Isolated Singular Solutions of Polynomial Systems: case of Breadth One”. In: *SIAM Journal on Numerical Analysis* 50.1 (2012), pp. 354–372.

[Lio40] J. Liouville. *Mémoire sur les transcendantes élliptiques de première et de seconde espèce, considérées comme fonctions de leur module*. 1840, pp. 441–464.

[Map18] Maplesoft. “Maple (2018)”. In: *a division of Waterloo Maple Inc., Waterloo, Ontario* (2018).

[MMM95] M. G. Marinari, T. Mora, and H. M. Möller. “Gröbner duality and multiplicities in polynomial system solving”. In: *Proceedings of the 1995 international symposium on Symbolic and algebraic computation*. 1995, pp. 167–179.

[Mez10] M. Mezzarobba. “NumGfun: a package for numerical and analytic computation with D-finite functions”. In: *Proceedings of the 2010 International Symposium on Symbolic and Algebraic Computation*. ACM. 2010, pp. 139–145.

[Mez16] M. Mezzarobba. *Rigorous Multiple-Precision Evaluation of D-Finite Functions in SageMath*. Tech. rep. 1607.01967. arXiv, 2016.

[MS10] M. Mezzarobba and B. Salvy. “Effective bounds for P-recursive sequences”. In: *Journal of Symbolic Computation* 45.10 (2010), pp. 1075–1096.

[MP98] M. Mezzino and M. Pinsky. “Leibniz’s Formula, Cauchy Majorants, and Linear Differential Equations”. In: *Mathematics magazine* 71.5 (1998), pp. 360–368.

[MKC09] R. E. Moore, R. B. Kearfott, and M. J. Cloud. *Introduction to interval analysis*. Vol. 110. Siam, 2009.

[RR84] H. Ratschek and J. Rokne. *Computer Methods for the Range of Functions*. Ellis Horwood Limited, 1984.

[SS00] M. Shub and S. Smale. “Complexity of Bezout’s theorem. I: geometric aspects”. In: (2000), pp. 1359–1401.

[Sma86] S. Smale. “Newton’s method estimates from data at one point”. In: *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics* (1986).

[SW05] A. J. Sommese and C. W. Wampler II. *The numerical solution of systems of polynomials*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005, pp. xxii+401. ISBN: 981-256-184-6.

[Sta78] R. P. Stanley. "Hilbert functions of graded algebras". In: *Advances in Mathematics* 28.1 (1978), pp. 57–83.

[Stu02] B. Sturmfels. *Solving systems of polynomial equations*. Number 97 in CBMS Reginal Conference Series in Mathematics. American Mathematical Soc., 2002.

[The18] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.3)*. `http://www.sagemath.org`. 2018.

[Ver99] J. Verschelde. "Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation". In: *ACM Trans. Math. Softw.* 25.2 (1999), pp. 251–276.

[Wol] Research Inc. Wolfram. *Mathematica, Version 11.3*. Champaign, IL, 2018.